Cotranslational protein assembly imposes evolutionary constraints on homomeric proteins

Cotranslational protein assembly imposes evolutionary constraints on homomeric proteins

Eviatar Natan[1]*, Tamaki Endoh[2], Liora Haim-Vilmovsky[3,4], Tilman Flock[5], Guilhem Chalancon[5], Jonathan TS. Hopper[6], Bálint Kintses[7], Peter Horvath[7,8], Lejla Daruka[7], Gergely Fekete[7], Csaba Pál[7], Balázs Papp[7], Erika Őszi[9], Zoltán Magyar[9], Joseph A. Marsh[10], Adrian H. Elcock[11], M Madan Babu[5], Carol V. Robinson[12], Naoki Sugimoto[2,13], Sarah A. Teichmann[4,14]*

[1]**The Aleph Lab Ltd, Oxford, OX2 8NU, UK**
[2]**Frontier Institute for Biomolecular Engineering Research (FIBER), Konan University, 7-1-20 Minatojimaminamimachi, Kobe, 650-0047, Japan**
[3]**EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK**
[4]**Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK**
[5]**MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 OQH, UK**
[6]**OMass Technologies Ltd, Centre for Innovation & Enterprise, Begbroke Science Park, Woodstock Road, Oxford OX5 1PF, UK**
[7]**Synthetic and System Biology Unit, Biological Research Center of the Hungarian Academia of Sciences, H-6726 Szeged, Hungary**
[8]**Institute for Molecular Medicine Finland, University of Helsinki, FI-00014, Helsinki**
[9]**Institute of Plant Biology, Biological Research Center of the Hungarian Academia of Sciences, H-6726 Szeged, Hungary**
[10]**MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK**
[11]**Department of Biochemistry, University of Iowa, Bowen Science Building, 51 Newton Road, Iowa City, Iowa 52242, USA**
[12]**Department of Chemistry, University of Oxford, 12 Mansfield Rd, Oxford, OX1 3TA, UK**
[13]**Graduate School of Frontiers of Innovative Research in Science and Technology (FIRST), Konan University, 7-1-20 Minatojimaminamimachi, Kobe, 650-0047, Japan.**
[14]**Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge CB3 0EH, UK**

***Co-corresponding authors**
Eviatar Natan: eviatarhj@gmail.com
Sarah A Teichmann: st9@sanger.ac.uk

Abstract

Cotranslational protein folding can facilitate rapid formation of functional structures. However, it could also precipitate premature assembly of protein complexes, if two interacting nascent chains are in close proximity. Through an analysis of protein structures, we show that homomeric contacts are enriched towards the C-termini of polypeptide chains across proteomes, which we hypothesize is the result of evolutionary selection for folding to occur prior to assembly. Using high-throughput imaging of homomers *in vivo* in *E. coli*, as well as engineered constructs with N- and C-terminal oligomerization domains, we show that proteins with C-terminal homomeric interface residues indeed consistently assemble more efficiently than those with N-terminal interface residues. With *in vivo*, *in vitro* and *in silico* experiments, we identify features of protein sequence and cellular environment that govern successful assembly of homomers, which have implications for protein design and expression optimization.

Introduction

Early in protein synthesis, a nascent chain begins to sprout from the ribosome's exit tunnel into the crowded cytoplasm. Some nascent chains will then fold concomitantly with translation, a process known as cotranslational folding[1-3]. Cotranslational folding is thought to have evolved to protect nascent chains from non-specific interactions with folded proteins, or from entanglement with other nascent chains.

While cotranslational folding, or folding proximal to the process of translation, can protect proteins from aggregation, it may also harbor a risk for homomers, which are protein complexes comprised of multiple identical subunits. Homomers are extremely common in all organisms, particularly prokaryotes, and are involved in all major cellular functions[4,5]. If a homomeric nascent chain folds during or soon after translation, it may also assemble[6,7]. We have coined the term *translational milieu* to describe the environment around a transcript that is being translated by ribosomes. Assembly of a nascent chain can occur with (*i*) a neighboring nascent chain being translated from the same mRNA, (*ii*) a proximal, mature subunit that was recently released from the same mRNA, or (*iii*) a nascent chain, or (*iv*) a mature protein that was translated by another copy of the mRNA in the same region of the cell[6] (Figure 1).

All of these scenarios constitute homomer assembly in the translational milieu. Importantly, for all these scenarios, misassembly could occur if assembly forces unfolded, partially folded or freshly folded parts of the polypeptide into extremely close proximity. This means that partially folded protein segments have an increased likelihood of interacting in a non-specific manner, thus increasing the chance of protein misassembly and aggregation.

Multi-domain proteins with repeats of domains of high sequence similarity have previously been shown to misfold through native-type interactions across different chains, resulting in a "domain-swap" scenario of chain entanglement and misfolding[8]. In these multi-domain proteins, the close proximity of similar domains increases the risk of misassembly, thus exerting evolutionary selection pressure[9]. The importance of sufficient folding time prior to the exposure of the unfolded nascent chain to the cellular environment

has been well-studied, both as a function of translation rate[10,11] and the requirement of folding of one domain prior to its exposure to the next translated domain[10].

For homomeric proteins, in addition to the synchronization between folding and translation, there is the constraint of coordinating assembly. Therefore, we hypothesize that evolutionary constraints must exist to ensure sufficient folding time prior to assembly. Specifically, if the position of interface-forming residues is such that they are translated first, *i.e.* at the N-terminal regions, this could allow incompletely translated chains to begin to assemble. If interface-forming residues are C-terminal, it is more likely to promote efficient assembly, as translation and folding of the majority of the protein would be completed prior to initiation of assembly.

Results

Homomeric but not heteromeric interface-residues have a C-terminal bias

The assembly of a complex is strongly dependent on the availability of its subunits. Once the residues that participate in assembly fold, whether co- or post-translationally, assembly can occur. However, such an assembly will force other partially unfolded parts of the protein to be in close proximity, increasing the risk of misassembly. To allow sufficient time for folding prior to assembly, we hypothesized that one way of minimizing the potential for misassembly would be to localize interface residues towards the C-terminus of the polypeptide chain. Therefore, we examined the locations of homomeric and heteromeric interface residues in a large set of non-redundant structures.

Our calculations showed a relative enrichment of residues forming homomeric interfaces from N- to C-termini, considering all proteins in our dataset (see Supplementary Methods, "Structural analysis of interface location"). Strikingly, there is a highly significant tendency for interfaces to be formed by residues in the C-terminal halves of proteins (Figure S1). Overall, there is an 11.3% greater chance that a given point on a protein's surface will be involved in a homomeric interface if it is located on the C-terminal half of the protein relative to the N-terminal half. This trend is also conserved across evolution (Figure 2A-B), and outliers in either direction can be explained by the small datasets for these species (*i.e.* P. horikoshii and R. norvegicus, Figure S1F). Finally, when bacterial or eukaryotic complexes are considered collectively, a significant enrichment in C-terminal interface residues is also conserved for each group.

The C-terminal enrichment of interface residues holds across different types of homomers. When we group proteins on the basis of their length, the significant C-terminal enrichment is observed across short and long proteins (Figure S2A). Moreover, we also observe significant C-terminal enrichments for homomers belonging to different symmetry groups. In contrast, for homomers with asymmetric structures, which we know are mostly the result of quaternary structure assignment errors and thus likely to have non-biological interfaces[19], there is no C-terminal interface enrichment (Figure S2A). Overall, this consistent enrichment of interface-forming residues in the C-terminal

halves of homomers provides strong evidence of selection pressure on this evolutionary feature of protein sequences.

It is also interesting to note that the interface enrichment from N- to C-termini is not completely uniform. Instead, there are two peaks of interface enrichment, centered at approximately 0.65 and 0.95. The precise origin of this is unclear as it is seen in different evolutionary groups and in proteins of different lengths. However, it is notable that this two-peak trend is much stronger in homomers with larger dihedral and cyclic symmetries, as compared to C_2 symmetric dimers. This suggests that it might be due to dihedral and cyclic homomers requiring at least two distinct interface-forming surface patches to assemble, whereas C_2 dimers require only a single surface[19].

Our measure of interface enrichment is already normalized for the amount of surface area exposed from N-to-C termini. However, to clarify this even further we have added a new Figure S3B, in which the relative interface enrichment and surface enrichment are presented separately. This shows that there is some tendency for proteins to expose more surface area towards their C-termini, while residues near the N-terminus are more likely to be buried. However, the interface enrichment is much stronger than the surface area trend..

The local concentration of homomeric proteins in the translational milieu is higher for homomers than heteromers on average, due to polyribosomes in both eukaryotes and prokaryotes, and co-transcriptional translation in bacteria. Therefore, there should be a lower propensity for premature assembly in heteromers, leading us to predict a weaker bias toward C-terminal interface enrichment for heteromers. Indeed, there is only a slight 1.6% interface enrichment in the C-terminal halves of heteromers, which is far weaker than for homomers and not statistically significantly (Figure S1C). Moreover, when split up on the basis of species or evolutionary group, the results are not consistent (Figure S1D-E). Bacteria have a slight interface-enrichment in their N-terminal halves, while eukaryotes and archaea have insignificant enrichments in their C-terminal halves. This suggests that the interface enrichment we observe relates to an evolutionary pressure specific to homomers.

I*n vivo* screen confirms increased misassembly of homomers with N-terminal interface enrichment

We carried out an *in vivo* image-based high-throughput screen with a set of 611 native *E. coli* homomers (Figure 2C-E). A flowchart summarizing the methodology is in Figure S3A. Briefly, we over-expressed a C-terminal GFP-fused version of each homomer[20], and applied a supervised machine-learning approach to automatically analyze the images of approximately one thousand cells per protein. The intensity of the fluorescence signal reflects the stability of the homomer[21]. For simplicity, cells were assigned to one of two groups: cells with homogeneous GFP signal throughout the cell, which we will refer to as 'Green' cells, and cells without GFP signal, which we will refer as 'Dark' cells. While 'Green' cells indicate a folded and soluble protein, 'Dark' cells may indicate on one of two scenarios: (i) the protein has an expression level that is below the detection limit or (ii) the protein aggregates prior to proper folding, as GFP folds cotranslationally[22]. By performing Western Blot analysis, we excluded 'Dark' homomers with low expression and only 'Dark' homomers were retained that have comparable expression levels to 'Green' ones (Figure 2E, Supplementary Data Sets 1 and 2). These procedures resulted in 203 'Green' and 109 'Dark' homomers.

Next, we asked whether high aggregation tendency correlates with N-terminal enrichment of homomeric interface residues, which would support the above analysis of protein structures. Indeed, the homomers that were associated with 'Dark' cells were significantly enriched in N-terminal interface residues as compared to the soluble homomers, *i.e.* 'Green' cells, which have more interface residues in the C-terminal halves of the chains (Figure 2).

When homomers were grouped on the basis of their length or relative interface size (interface-size/total surface-size), the enrichment of interface residues in the N-terminal half remained evident for 'Dark' homomers as compared to 'Green' ones across all categories (Figure S3B and S3C, respectively). Thus, in line with the results of the structural analysis, neither length-dependent folding rate nor relative interface size appear to be major determinants of this phenomenon. Last, we retained only homomers with

cytoplasmic cellular localization (Figure S3D). The relative enrichment of N-terminal interface residues in the 'Dark' group was still present.

The position of the oligomerization domain determines assembly and solubility in engineered constructs

A possible explanation for the C-terminal preference for interface residues is selection against detrimental premature assembly, or misassembly, as indicated by the *in vivo* screen. The synthesis of the interface early in translation, *i.e.* at the N- rather than C-terminus, increases the propensity for assembly to occur during, or soon after translation. Although such early assembly may be beneficial for some proteins[6,23], it can also lead to misassembly due to an increase in nonspecific interactions between partially unfolded nascent chains (Figure 1).

To test this and dissect the underlying mechanism of the observed N-*versus* C-terminal bias, we built a library of constructs that reflects the different characteristics of homomers, where each construct is comprised of three components (See also Table S1, Supplementary Note 1 and Figure S4A): (i) A short amino acids oligomerization domain (Tet) placed at either the N- or C-terminus of the constructs. Conveniently, a single amino-acid substitution in this small domain determines whether the domain forms tetramers, dimers, or remains monomeric in its folded state[14,15]

Homomers must often assemble very close to their translation environment due to macromolecular crowding, which limits diffusion rates, and the Tet domain is even more likely to do so. First, the Tet oligomerization domain has a low (~nM) dissociation constant[24], which supports *in vivo* observations that Tet exists as oligomers[14]. Second, Tet folds and assembles faster than its translation rate[12], which means it is expected to fold soon after it exits the ribosome tunnel, folding cotranslationally as observed previously[13]. Therefore, if positioned at the N-terminus, Tet is likely to oligomerize during translation of the reporter domain, *i.e.* cotranslationally. However, if positioned at the C-terminus, the short Tet domain can only assemble after leaving the ribosome exit tunnel, constituting the last amino acids to be translated.

Using ESI-MS, we confirm that the tetrameric constructs can generate tetrameric quaternary structures, and the monomeric consists of a single subunit (Figure S4E-H).

(ii) The second component is the reporter domain, which was chosen based on a detectable signal and folding rate. The reporters, YFP, two versions of GFP and Luciferase, are each comprised of a few hundred amino acids. (Sequences provided in Supplementary Methods). Thus, their translation is orders of magnitude longer than the Tet folding-rate[12,16,17], considering that bacterial translation rates are on the order of 10-20 amino acids per second[18].

(iii) The third component of the constructs is a linker separating the Tet oligomerization domain and the reporter. The linker was designed to be flexible and of diverse lengths. The length effectively controls the local concentration of the oligomerization domain, because the linker enforces a spatial separation between any oligomerization domains emanating from ribosomes on the same mRNA molecule. Three different linkers were used: a short-linker (SL), which is a five amino acid (aa), glycine-based linker. The medium-linker (ML) and long-linker (LL) are comprised of the lipoyl domains of the dihydrolipoyl acetyltransferase enzyme[27]. The medium linker (ML) is 50aa long, which is the lipoyl domain from *B. stearothermophilus*, and the long linker (LL) is 100aa long, which also contains the lipoyl domain from *E. coli*[28,29]. The library is divided into sub-libraries according to the reporter domain and the linker length (Table S1).

We first investigated two constructs with identical domain composition: an oligomerization domain (Tet) connected to a reporter (YFP) by a short linker (SL) on either its N- or C-terminus. Both constructs form the same tetrameric quaternary structure (Figure S4E-G). Interestingly, and in agreement with the analysis of the bioinformatics and high-throughput analysis in *E. coli*assays, we observed, a significant difference in fluorescence levels between the two constructs using confocal microscopy (Figure 3A). This is a remarkable difference considering the similarities of the proteins in sequence composition and quaternary structure (Figure S4). Importantly, expression levels were very similar based on Western blotting (Figure 3B and S4D). However, and in agreement with the microscopy data, a clear difference was

observed in protein solubility,: only YFP-SL-Tet showed a band in the soluble fraction, indicating a folded state, while Tet-SL-YFP was found in the insoluble fraction. We can thus conclude that the difference in fluorescence is due to post-translational events.

Quantifying the above effect using flow cytometry, we found that the fluorescence is over an order of magnitude higher for the construct with the tetramerization domain at the C-terminus *versus* the N-terminus. This is in the same range as the difference observed in the *in vivo* high-throughput screen (Figure 2). Using a mutated, dimeric construct, which cannot form tetramers[14,15], we observed the same C- *versus* N-terminal fluorescence increase as for the tetrameric variant (Figure S4B-C), suggesting that the misassembly rates in these constructs are a function of assembly *per se* rather than a specific oligomeric state.

Interestingly, co-expression with the Tet-peptide significantly reduces misassembly (Figure 3E). The expression of Tet-SL-YFP, with or without the peptide, was normalized to a monomeric variant that does or does not co-express the peptide, respectively. The rapid association kinetics of the tetrameric variant[24] and its high expression[25] (not the case for the p53 full-length protein[26]) can explain this rescue, likely by masking the homomeric interface of the Tet-SL-YFP polypeptide by the Tet-peptide to prevent misassembly. In summary, analysis of expression and correct assembly of engineered constructs support a crucial role for the position of the oligomerization domain at the N- *versus* C-terminus.

Extending linker length reduces misassembly

We next sought to assess the effect of a long and flexible linker to the fate of the protein's stability. Three different linkers were used as described above.

For each sub-library, a monomeric variant was generated as a control by introducing a single point mutation in the Tet sequence. Thus we were able to calculate the ratio of fluorescence intensity between the tetrameric *versus* monomeric variants to quantify the contribution of homomerization to misassembly (Figure 4A).

Importantly, the three monomeric constructs, which differ in the length of their linker, showed a similar level of fluorescence (Figure S5). In contrast, the fluorescence intensity of the tetrameric constructs increases with increasing linker length. The ratio of the fluorescence of the tetramer-to-monomer strains of each linker showed a positive correlation between the length of the linker and the extent of correct assembly of the protein (Figure 4B). These results suggest that the increase in linker length is proportional to successful assembly rates. This could be as a result of the distance between the domains, or because the linker buffers non-specific interactions between the oligomerization and reporter domains, as the linker is soluble and globular.

Fast folding of the reporter promotes efficient assembly

The balance between translation and folding rate is crucial for the fate of synthesized proteins. For example, changes in translation rates *via* a small number[3] or even a single[30] synonymous substitution of a rare to abundant tRNA codon changes the translation-folding balance and affects folding efficiency[31]. This is because slower translation rates provide a longer co-translational folding time[32].

To examine the role of protein folding rate in misassembly, we used two monomeric GFP variants with different folding rates: a fast folding GFP variant (fGFP)[33] and a wild-type variant (GFP) with a slow folding rate[34]. In order to isolate the effect of folding rate, we used a long linker, as it maintains the same fluorescence level for both monomeric and tetrameric variants of the fast folding YFP (Figure 4). As expected, using confocal microscopy, the monomeric and tetrameric fGFP variants presented similar fluorescence levels (Figure 4C-D). This similarity to the YFP results is not surprising, as fGFP and YFP share the three fast folding mutations (F64L, V68L, S72A) located at the center of the beta-barrel[33].

In contrast, the slow folding GFP (GFP) showed a significant difference between the monomeric and tetrameric variants. To quantify these observations, we used flow cytometry under the same experimental conditions. While the monomeric and tetrameric fGFP variants have essentially the same fluorescence levels, the tetrameric GFP has ~3.5-fold lower fluorescence than

the monomeric GFP variant (Figure 4D). Interestingly, by culturing the cells with the tetrameric and monomeric GFP variants at 18°C, the observed difference was reduced significantly (Figure S5).

Luciferase (Luc) is a long, two-domain and slow folding protein, with a completely different architecture to the beta-barrel fluorescent proteins. We cloned a Luciferase sub-library with both short- and long linkers (Table S1). Similar to the trend observed for the other reporter genes, the tetrameric Luc variant with either a short- or long-linker, had a lower luminescence level, which indicates a higher misassembly rate compared to the monomeric Luc variants (Figure S6).

Misassembly and recovery using an *in vitro* translation system to tune mRNA:ribosome ratio

To further investigate whether this phenomenon occurs cotranslationally or soon after, *i.e.* in the *translational milieu*, we expressed the constructs in the PURE *in vitro* translational system, which is a well characterized system that allows full control of all required components and their concentration[35]. For example, by increasing the [mRNA:ribosome] ratio, translation occurs under monosomic rather than polysomic conditions, thus decreasing the probability of nascent chains interacting with each other proximal to their translation sites.

We first examined monomeric *versus* tetrameric long linker GFP variants (Figure 5). The levels of correctly folded reporter translated at high ribosome density, *i.e.* low [mRNA:ribosome] ratio, were quantified as the ratio of [fluorescence to protein expression level] (Figure S7). The tetrameric variant had ~6-fold lower fluorescence level than its monomeric counterpart (Figure 5A), which is in agreement with the *in vivo* results. Moreover, for the fast folding fGFP reporter, the difference between the monomeric and tetrameric variants was much lower, with only ~2-fold difference. To examine the generality of these findings, we investigated the Luc reporter both *in vivo* and *in vitro* (Figure S6). The tetrameric variant with either short- or long-linker, had a lower luminescence level, which indicates a higher misassembly rate compared to the monomeric variants.

By decreasing the [mRNA:ribosome] ratio by 150-fold, the probability of polyribosome formation is drastically reduced. Therefore, the local concentration of the nascent chains, and consequently their probability to assemble during or soon after translation, significantly decreases. It is worth mentioning that both polysomic and monosomic reactions had a similar total expression level at the time that the measurements were taken (Figure S7). This is due the fact that a sufficient time was given for both reactions to reach saturation (See also Supplementary Methods). Therefore, there are two major differences between the two conditions, that both affect the *translation milieu*: the proximity between translating nascent chains, and the accumulation rate of the translated protein as it increases in the *translation milieu*.

The results of the fGFP and Luciferase sub-libraries under monosomic and polysomic conditions also support the proposed hypothesis. For example, the low local concentration of nascent chains, as in the monosomic condition, rescues misassembly of the slow folding Luc reporter (Figure 5B and Figure S6). Moreover, and in contrast to Luc, the fast folding fGFP tetrameric variant showed only a marginal difference between the two conditions, and no difference was observed for its monomeric variants.

Some chaperones reduce *in vitro* misassembly

The selectivity of the PURE system allows us to test the effect of different chaperones groups on rescuing constructs from misassembly (Figure 5 and Figure S8). We tested three chaperone groups. The first includes DnaK, DnaJ and GrpE, (KJE-mix). The second includes GroEL and GroES, namely GroE-mix, and the third is Trigger Factor (TF). The TF ribosome-associated chaperone interacts directly with unfolded nascent polypeptide chains as they emerge from the ribosome exit tunnel[36], allowing small domains to fold under its "cradle". It has been shown to have little effect on rescue of cotranslational misassembly[7]. This is in full agreement with our results as TF showed no significant effect (Figure S8).

On the other hand, the KJE-mix had the largest effect of all chaperone groups, an effect that correlates with the proteins' oligomeric state. Interestingly, the overall profile of the chaperones correlated with the proteins' folding rate of

the reporter rather than their fold similarities. For example, the effect on the slow folding tetrameric protein GFP is more similar to the tetrameric Luc rather than to fGFP, with which it shares >95% sequence similarity. Last, the GroEL mix had an effect only on the tetrameric variants with relatively slow folding, *i.e.* GFP and Luc, but not their corresponding monomeric variants. This is in agreement with previous work showing that reactivation of (monomeric) Luc was observed with a KJE-mix, but not with Gro-mix[37,38].

KJE should interact with GFP and Luc to aid their folding, *e.g.* Ref[39]. From our data, we cannot tell whether the chaperones interact with the constructs prior to the translation of GFP or Luc. Examining high throughput data of previous work in *E. coli*[40], we investigated interaction enrichment of homomers and heteromers with chaperones. We could find a significant number of *E. coli* homomeric complexes interacting with chaperones, although no significant difference is detected between homomeric and heteromeric complexes (Figure S8). Nevertheless, it is clear that these chaperones reduce the overall misassembly level, which provides some explanation for why homomeric contacts are tolerated in N-terminal positions in naturally occurring proteins, albeit at a lower rate than expected by chance.

In silico **simulations visualize cotranslational assembly**

To estimate the probability of nascent chain interactions occurring in the context of polyribosomes, and to gain insight into the mechanism of cotranslational assembly at atomic detail, we carried out *in silico* simulations of translation, folding and assembly. We used coarse-grained residue-level Brownian-dynamics simulations for three representative constructs with the YFP reporter. We focused on an inter-ribosomal geometry identified in tomographic reconstructions of experimentally determined *E. coli* ribosomes[41], which has two peptide exit tunnels in close proximity. Using this model, we observed cotranslational folding and assembly, posttranslational assembly, and simulations where no assembly occurs (Figure 6 and Videos S1-3, doi: https://figshare.com/s/48830eb9d72a80065d4a).

We found that in simulations of the ribosomal synthesis of tetrameric N-terminal constructs (Tet-SL-YFP), cotranslational assembly of the constructs

occurred in 90% of the simulations. As a result of this high-frequency cotranslational assembly, intermolecular interactions of the YFP domains occurred in 75% of the simulations (see also Figure S9). These intermolecular interactions likely represent misassembly events that inhibit the development of the fluorescence of the naturally monomeric YFP domain.

Extending the linker connecting the Tet and the YFP, decreases misassembly-like events. The reason was not due to a decrease in cotranslational assembly events mediated by the Tet domains, which are similar to the short linker construct, but due to fewer YFP-YFP interactions. These results correlated well with the *in vivo* and *in vitro* results, again highlighting the ameliorating role of the long linker between the oligomerization domain and reporter domains, as a diluter of the local concentration of the domains.

The simulations of tetrameric C-terminal construct (YFP-SL-Tet) showed much less frequent cotranslational assembly events, in agreement with our experimental results. When assembly did occur, it was a posttranslational event, or occurred as the newly synthesized chains were in the process of diffusing away from the ribosome exit tunnels. As a consequence, and as expected from our hypothesis, intermolecular interactions of YFP to YFP were rare events for the C-terminal construct.

These results indicate a clear relationship between the positioning of the Tet domain and the likelihood of misassembly events preventing the reporter domain's fluorescence.

Homomer misassembly reduces fitness and mediate negative selection

To assess the degree of homomer misassembly on the global fitness of *E. coli*, we measured the real-time growth-rates of strains expressing the YFP, fGFP and GFP sub-libraries and compared strains with N- or C-terminal constructs. In agreement with the other approaches used in this work, we found that YFP showed no significant difference between the growth rates of the monomeric and @C-tetrameric variants, which are both different to the @N-tetrameric variants (Figure S10). A significant trend in favor of monomeric over @N-tetrameric variants was also observed for fGFP and GFP.

463 Under similar settings, the two YFP tetrameric constructs were expressed
464 in *E. coli* and examined using complex immuno-precipitation and proteomics
465 characterization experiments. We found a five-fold increase in the chaperone
466 HtpG, a bacterial homologue of Hsp90, for the tetrameric N-terminal construct
467 (Supplementary Data Set 3). These *in vivo* results suggest that misassembly
468 represents a burden to the cell that has a direct effect on growth rate and thus
469 cellular fitness. This selection pressure will ultimately affect the position of
470 residues involved in homomerization.
471
472

Discussion

A protein's amino acid sequence determines its structure, stability and interactions with other biomolecules. For homomeric proteins, which dominate protein quaternary structure space, the relationship between these parameters is only partially understood[4,6,42]. The stability of the monomer, meaning its capacity to maintain the correct fold, is crucial for the stability of the entire complex. However, assembly, *i.e*. the protein's native interactions with another identical chain to form a homomeric complex, can occur once the interface residues are available after translation and folding[43]. It is therefore plausible that assembly will take place prematurely, leading to incorrect assembly (misassembly), imposing a burden on the cell and thus decreasing cellular fitness. Here, we hypothesized that separation between the synthesis and assembly must be ensured to guarantee a complex's stability.

One way of achieving this is to position the residues mediating the assembly towards the end of the protein, so that it is synthesized before it starts assembling. Interestingly, it has been previously shown *in vitro* that refolding after denaturation of homomeric proteins is more challenging than for monomeric proteins potentially due to misassembly[44]. This suggests that ribosomal protein synthesis may actually play a role in fine-tuning the correct assembly of homomers.

To further explore whether interface location and linker length are important for correct assembly of native proteins, we searched for *E. coli* homomers with full-length crystal structures, well-defined oligomerization domains and predicted post- or cotranslational folding signatures as calculated by O'Brien *et al.*[31]. We identified three such proteins meeting these criteria (Figure S10 and Supplementary Data Set 4). Two of the *E. coli* homomers have oligomerization domains located safely towards the C-termini, thus avoiding premature (mis)assembly. However, one of these three structures has an oligomerization domain at the N-terminus. In agreement with our prediction, the protein has a long linker right after the oligomerization domain. Moreover, the oligomerization domain is also predicted to fold posttranslationally[33], which can provide an additional protection via temporal separation of folding and assembly, *i.e.*, late assembly, to avoid misassembly.

While our work identified several other factors of the cellular environment such as the role of chaperones or and ribosome density as a countermeasure strategy to cope with homomeric misassembly, a more efficient approach would be to avoid premature assembly in the first place. In other words, evolving protein sequences to ensure a correct balance between translation, folding and assembly. The remarkable consistent results of these analyses throughout our work allow us to put forward a spatiotemporal framework that strongly supports such a primary mechanism. These factors are summarized in Figure 7.

Importantly, we would like to emphasize that many other factors, including some that were examined in this work, must be involved in the critical mechanism of protein assembly. These may include the secondary structure of mRNA, ribosome density, mRNA local concentration, overall protein translation rate, assembly interface and affinity, and the aggregation propensity of each domain. We encourage others to explore these factors, as well as similarities between homomers and heteromers of bacterial operons.

Interactions between polypeptide chains are inter-molecular, stochastic events where the frequency and length of association are determined by the nature of the protein's surface[43,45]. Therefore, in the confined environment of the translational milieu, where assembly may be an intra-molecular event competing with the intra-molecular folding, a single mutation can have a greater effect than anticipated. For example, a mutation that even weakly promotes a steady or transient interaction may have a significant effect on the stability of the protein. Moreover, our work may also explain directionality of truncation in circular permutation constructs[46].

Importantly, we now hypothesize that misassembly in the *translation milieu* contributes toward diseases such as the neurodegenerative Huntington's disease (HD), via misassembly of the Huntington protein. The short N-terminal domain of the Huntington protein promotes oligomerization, and consequently significantly accelerates amyloid-formation[47]. It is therefore tempting to speculate that oligomerization, and thus amyloid formation, occurs in the *translation milieu*, which suggests new strategies for tackling this disease.

Figure Legends

Figure 1. Illustration of the possible cotranslational assembly of homomeric proteins.
The *translation milieu*, *i.e.* the immediate environment around the mRNA, is enriched by the translated nascent chain and protein, which for homomers immediately increases the propensity for oligomerization. Oligomerization of homomers at the *translation milieu* can occur cotranslationally, *i.e.* between nascent chains of the same mRNA or between nascent chains of identical mRNAs copies. Alternatively oligomerization can occur between a nascent chain and a fully-translated protein. In either scenario, premature assembly (or misassembly) of partially folded proteins can occur.

Figure 2. Interface residues of native homomers are C-terminally enriched, which correlates with *in vivo* stability of the protein. (A-B) Distribution of interface-forming residues in the N- vs. C-terminal halves of homomeric proteins. (A) Relative enrichment of interface-forming residues along the lengths of homomers, combing structures across all evolutionary groups. Residues are binned according to their position along the full-length protein (N-terminus is 0, the C-terminus is 1). Error bars represent standard error calculated from 10^6 bootstrapping replicates, and the *p*-value is derived from the fact that there was a net enrichment in C-terminal interface for all replicates. (B) Relative enrichment in interface in the C-terminal halves of proteins compared to the N-terminal halves for all species with >100 non-redundant homomer structures in our dataset. Error bars represent standard error calculated from 10^4 bootstrapping replicates *per* species. The number of homomer structures *per* species is given in parentheses. The non-redundant sets of homomeric and heteromeric complexes are provided in Supplementary Data Set 5. (C) Image-based high-throughput screen reveals N-terminal enrichment of interface residues in aggregating homomers. The relative enrichment of interface-forming residues along the protein length is shown in green and grey for 'Green' and 'Dark' cells, respectively. (**p-value <0.01, *p-value <0.05. Error bars represent *s.d.*). (D) The fluorescence of the 'Green' and 'Dark' cells groups is significantly different (with *p-value* of $2.2e^{-16}$, Wilcox rank test). Fluorescence 0 is equal to the mean fluorescence of the negative control, *i.e. E. coli* cells without GFP. (E) The expression level in the 'Dark' homomer group is similar to that of the 'Green' homomer group based on the Western Blot analysis (*p-value* from Mann-Whitney U-test 0.265) as presented in Supplementary Data Set 2.

Figure 3. *In vivo* expression of the constructs shows that the position of the oligomerization domain is crucial for the solubility of the expressed protein. (A) Confocal microscopy images of Tet-SL-YFP and YFP-SL-Tet, with YFP reporter gene. The latter presented homogenous fluorescence throughout the cell, in stark contrast to Tet-SL-YFP. (B) A Western Blot shows that the expression level of both constructs was similar, yet only YFP-SL-Tet was present in the soluble fraction, meaning the origin of the phenotype lies in a post-transcriptional event. The full WB gel appears in Figure S4D. (C) Fluorescence levels of each strain using flow Cytometry. (D) Mean fluorescence intensity as measured in (C) (E) Co-expression of the tetrameric (Tet@N) construct with a Tet-peptide decreased the level of misassembly. The results are the mean of tetrameric-to-monomeric variants ratio (these variants differ only by a single amino acid). On the left is the ratio of these variants without co-expression of Tet-peptide. On the right is the ratio of these variants, with both expressing Tet-peptide. [Independent cell culture replicates (n>5), **p-value <0.01, *p-value <0.05, double sided t-test. Error bars represent *s.d.*].

Figure 4. Extending the linker decreases misassembly rates.
(A) Scheme of the different constructs. All constructs have an oligomerization domain at the N-terminus. The comparison is between tetrameric and monomeric variants, which differ by a single amino acid. The medium- (ML), and long (LL) linkers are one, and two flexible lipoyl domains, respectively. On the right is an unscaled diagram of the nascent-chain ribosome complexes after the translation of the linker and before the reporter exit the ribosome tunnel. (B) **Flow cytometry** analysis provides us with the ratio of the fluorescence of tetrameric-to-monomeric variants, which is proportional to the length of the linker. This reason for this correlation could be associated with the increased distance between translated reporter polypeptides, or due to the presence of the soluble linker serving as a barrier between the translating reporters. (C) Confocal microscopy images of constructs with fGFP or GFP reporter genes. Similarly to the long-linker YFP variants, fGFP shows high fluorescence without a significant difference between the monomeric and tetrameric variants (no saturation was allowed). For the GFP variants, a strong effect is observed under the same conditions. (D) Flow cytometry analysis of the mean ratio of tetrameric-to-monomeric variants for each reporter gene. As seen in (C), no significant difference was observed between the fGFP variants. However a 3.5-fold higher fluorescence was observed for the GFP monomeric compared to tetrameric variants. [Flow cytometry plots are available in Figure S5. Independent cell culture replicates (n>5), ***p*-value <0.01, **p*-value <0.05, double sided t-test. Error bars represent *s.d.*].

Figure 5. Misassembly as a function of oligomerization, folding-rate and ribosome density, using PURE *in vitro* translation system.
(A) The fluorescence difference between the tetrameric and monomeric variants of fast folding (fGFP) and slow folding GFP. The value of the tetrameric construct was divided by the monomeric constructs of the same reporter gene and presented as average. (B) The fluorescence difference between high and low [mRNA:ribosome] ratio. The GFP tetrameric variants showed a decrease in solubility in comparison to the monomeric variant, a difference that is rescued if the conditions are shifted towards monosomic conditions. fGFP variants show the same solubility in both polysomic and monosomic conditions. *In vivo and in vitro* experiments were conducted using the slow folded Luciferase reporter gene, showing the exact same trends as for the slow folding GFP. (C) Fast and slow GFP folding reporters were tested using three chaperone groups, KJE-mix, GroE-mix, and Trigger Factor. Interestingly, only slow folding GFP was affected by chaperones, and only by the KJE-mix. (D) Summary of the effect of the different chaperones on GFP, fGFP and Luc sub-libraries. The data is presented as the ratio of signal in a sample with chaperones *vs.* expression in the absence of chaperones. Measurements were repeated at least three times and averaged (see Supplementary Methods). Overall, the effect of the KJE-mix chaperones correlated with oligomeric state, *i.e.* tetramer *versus* monomer and with folding-rate, *i.e.*, fast- and slow-folding proteins. The highest rescue effect was achieved for the tetrameric slow folding Luc and GFP. (*p-value* * < 0.05, ** < 0.01, NS = Not Significant, double sided t-test. Error bars represent *s.d.*).

Figure 6. *In silico* simulation of translation of different constructs.
(A) Simulation snapshot of cotranslational folding of two neighboring nascent chains of Tet-SL-YFP and YFP-SL-Tet. Composite plot showing regions typically sampled by the two nascent Tet-SL-YFP chains up to the point at which the translation of the first chain was completed. The leading ribosome is shown on the left. (B-C) Cotranslational events as captured by simulations of polysomic translation. The relative positioning of the two ribosomes as found previously[41]. The blue and pink over the ribosomes represent positively and negatively charged amino acids, respectively. In the case of Tet-SL-YFP the two chains intermingle, as the oligomerization domains assemble

cotranslationally. (C) Same as (B) but showing a typical result for the YFP-SL-Tet construct, that is no intermingling of the two nascent chains occurs because the oligomerization domains were not translated yet and thus remain unassembled. (D) Simulation snapshot showing the cotranslational assembly of two neighbouring nascent chains. The leading ribosome is shown on the left. Tet is in red and YFP in yellow. (E) Table showing the number of co- or post-translational (in brackets) assembly events, misassembly-like events and total number of simulations.

Figure 7. Summary scheme. Cotranslational (mis)assembly and consequences as a function of sequence-intrinsic features. (A) The propensity to assemble requires generation of a sufficiently folded interface. Then, influenced by the frequency and nature of the interface encounters, a successful assembly (right) or misassembly (left) occurs. For instance the oligomerization domain position, the length of the linker and folding rate of the reporter-domain are a few of the determining factors in this balance (Red circle mature protein). (B) The factors explored in this work that determine successful assembly. (C) Successful cotranslational assembly depends on the balance between the kinetics of translation, folding and assembly.

References

1. Elcock, A.H. Molecular simulations of cotranslational protein folding: fragment stabilities, folding cooperativity, and trapping in the ribosome. *PLoS Comput Biol* **2**, e98 (2006).
2. Sander, I.M., Chaney, J.L. & Clark, P.L. Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design. *J Am Chem Soc* **136**, 858-61 (2014).
3. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol* **20**, 237-43 (2013).
4. Levy, E.D. & Teichmann, S. Structural, evolutionary, and assembly principles of protein oligomerization. *Prog Mol Biol Transl Sci* **117**, 25-51 (2013).
5. Goodsell, D.S. & Olson, A.J. Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct* **29**, 105-53 (2000).
6. Natan, E., Wells, J.N., Teichmann, S.A. & Marsh, J.A. Regulation, evolution and consequences of cotranslational protein complex assembly. *Curr Opin Struct Biol* **42**, 90-97 (2017).
7. Shieh, Y.W. et al. Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science* **350**, 678-80 (2015).
8. Borgia, M.B. et al. Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* **474**, 662-5 (2011).
9. Wright, C.F., Teichmann, S.A., Clarke, J. & Dobson, C.M. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* **438**, 878-81 (2005).
10. Nissley, D.A. & O'Brien, E.P. Timing is everything: unifying codon translation rates and nascent proteome behavior. *J Am Chem Soc* **136**, 17892-8 (2014).
11. Buhr, F. et al. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol Cell* **61**, 341-51 (2016).
12. Mateu, M.G., Sanchez Del Pino, M.M. & Fersht, A.R. Mechanism of folding and assembly of a small tetrameric protein domain from tumor suppressor p53. *Nat Struct Biol* **6**, 191-8 (1999).
13. Nicholls, C.D., McLure, K.G., Shields, M.A. & Lee, P.W. Biogenesis of p53 involves cotranslational dimerization of monomers and posttranslational dimerization of dimers. Implications on the dominant negative effect. *J Biol Chem* **277**, 12937-45 (2002).
14. Gaglia, G., Guan, Y., Shah, J.V. & Lahav, G. Activation and control of p53 tetramerization in individual living cells. *Proc Natl Acad Sci U S A* **110**, 15497-501 (2013).
15. Lomax, M.E., Barnes, D.M., Hupp, T.R., Picksley, S.M. & Camplejohn, R.S. Characterization of p53 oligomerization domain mutations isolated from Li-Fraumeni and Li-Fraumeni like family members. *Oncogene* **17**, 643-9 (1998).
16. Mateu, M.G. & Fersht, A.R. Mutually compensatory mutations during evolution of the tetramerization domain of tumor suppressor p53 lead to

impaired hetero-oligomerization. *Proc Natl Acad Sci U S A* **96**, 3595-9 (1999).
17. Mateu, M.G. & Fersht, A.R. Nine hydrophobic side chains are key determinants of the thermodynamic stability and oligomerization status of tumour suppressor p53 tetramerization domain. *EMBO J* **17**, 2748-58 (1998).
18. Iwasaki, S. & Ingolia, N.T. PROTEIN TRANSLATION. Seeing translation. *Science* **352**, 1391-2 (2016).
19. Ahnert, S.E., Marsh, J.A., Hernandez, H., Robinson, C.V. & Teichmann, S.A. Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015).
20. Kitagawa, M. et al. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res* **12**, 291-9 (2005).
21. Waldo, G.S., Standish, B.M., Berendzen, J. & Terwilliger, T.C. Rapid protein-folding assay using green fluorescent protein. *Nat Biotechnol* **17**, 691-5 (1999).
22. Ugrinov, K.G. & Clark, P.L. Cotranslational folding increases GFP folding yield. *Biophys J* **98**, 1312-20 (2010).
23. Wells, J.N., Bergendahl, L.T. & Marsh, J.A. Co-translational assembly of protein complexes. *Biochem Soc Trans* **43**, 1221-6 (2015).
24. Rajagopalan, S., Huang, F. & Fersht, A.R. Single-Molecule characterization of oligomerization kinetics and equilibria of the tumor suppressor p53. *Nucleic Acids Res* **39**, 2294-303 (2011).
25. Natan, E. & Joerger, A.C. Structure and kinetic stability of the p63 tetramerization domain. *J Mol Biol* **415**, 503-13 (2012).
26. Natan, E. et al. Interaction of the p53 DNA-binding domain with its n-terminal extension modulates the stability of the p53 tetramer. *J Mol Biol* **409**, 358-68 (2011).
27. Jones, D.D., Stott, K.M., Howard, M.J. & Perham, R.N. Restricted motion of the lipoyl-lysine swinging arm in the pyruvate dehydrogenase complex of Escherichia coli. *Biochemistry* **39**, 8448-59 (2000).
28. Radford, S.E., Laue, E.D., Perham, R.N., Martin, S.R. & Appella, E. Conformational flexibility and folding of synthetic peptides representing an interdomain segment of polypeptide chain in the pyruvate dehydrogenase multienzyme complex of Escherichia coli. *J Biol Chem* **264**, 767-75 (1989).
29. Lengyel, J.S. et al. Extended polypeptide linkers establish the spatial architecture of a pyruvate dehydrogenase multienzyme complex. *Structure* **16**, 93-103 (2008).
30. Tsai, C.J. et al. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J Mol Biol* **383**, 281-91 (2008).
31. O'Brien, E.P., Vendruscolo, M. & Dobson, C.M. Prediction of variable translation rate effects on cotranslational protein folding. *Nat Commun* **3**, 868 (2012).
32. Zhang, G. & Ignatova, Z. Folding at the birth of the nascent chain: coordinating translation with co-translational folding. *Curr Opin Struct Biol* **21**, 25-31 (2011).

33. Xu, C., Wang, S., Thibault, G. & Ng, D.T. Futile protein folding cycles in the ER are terminated by the unfolded protein O-mannosylation pathway. *Science* **340**, 978-81 (2013).
34. Reid, B.G. & Flynn, G.C. Chromophore formation in green fluorescent protein. *Biochemistry* **36**, 6786-91 (1997).
35. Shimizu, Y., Kanamori, T. & Ueda, T. Protein synthesis by pure translation systems. *Methods* **36**, 299-304 (2005).
36. O'Brien, E.P., Christodoulou, J., Vendruscolo, M. & Dobson, C.M. Trigger factor slows co-translational folding through kinetic trapping while sterically protecting the nascent chain from aberrant cytosolic interactions. *J Am Chem Soc* **134**, 10920-32 (2012).
37. Niwa, T., Kanamori, T., Ueda, T. & Taguchi, H. Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proc Natl Acad Sci U S A* **109**, 8937-42 (2012).
38. Jaenicke, R. Protein folding: local structures, domains, subunits, and assemblies. *Biochemistry* **30**, 3147-61 (1991).
39. Schroder, H., Langer, T., Hartl, F.U. & Bukau, B. DnaK, DnaJ and GrpE form a cellular chaperone machinery capable of repairing heat-induced protein damage. *EMBO J* **12**, 4137-44 (1993).
40. Calloni, G. et al. DnaK functions as a central hub in the E. coli chaperone network. *Cell Rep* **1**, 251-64 (2012).
41. Brandt, F. et al. The native 3D organization of bacterial polysomes. *Cell* **136**, 261-71 (2009).
42. Marsh, J.A. & Teichmann, S.A. Structure, Dynamics, Assembly, and Evolution of Protein Complexes. *Annu Rev Biochem* (2014).
43. Levy, E.D., De, S. & Teichmann, S.A. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc Natl Acad Sci U S A* **109**, 20461-6 (2012).
44. Jaenicke, R. & Lilie, H. Folding and association of oligomeric and multimeric proteins. *Adv Protein Chem* **53**, 329-401 (2000).
45. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E.D. Proteins evolve on the edge of supramolecular self-assembly. *Nature* (2017).
46. Peisajovich, S.G., Rockah, L. & Tawfik, D.S. Evolution of new protein topologies through multistep gene rearrangements. *Nat Genet* **38**, 168-74 (2006).
47. Tam, S. et al. The chaperonin TRiC blocks a huntingtin sequence element that promotes the conformational switch to aggregation. *Nat Struct Mol Biol* **16**, 1279-85 (2009).

Online Methods

Structural analysis of interface location. The entire set of X-ray crystal structures was taken from the Protein Data Bank (PDB) on 2015-03-19. Only protein chains >30 residues in length were considered, and structures with >10% non-protein heavy atoms (ignoring water) were excluded. Structures with known quaternary structure assignment errors[48] were excluded. Structures were then filtered for sequence redundancy at the level of 50% sequence identity, as previously described[49]. Two non-redundant datasets were generated: i) redundancy filtering was performed across all structures; ii) redundancy filtering was performed only for members of the same species, for the species-specific dataset in Figure 2B. The non-redundant sets of homomeric and heteromeric complexes are provided in Supplementary Data Set 5.

Residue-specific solvent accessible surface area was calculated as in Ref[50]. The amount of interface formed by a residue was taken as the differences between its accessible surface area as a monomer and its accessible surface area within the complex. Each structured residue within a PDB structure was then mapped back to its position in the corresponding Uniprot sequence (taken from the PDB *db_id* field), and its relative position within the full-length protein was used for all analyses. To control for the fact that residues along the length of a polypeptide chain are not all equally likely to occur at the protein surface and thus form interface (see Figure S2C), the interface enrichment is normalized by the overall distribution of solvent exposed residues in the monomeric subunits. Finally, the C-terminal enrichment indicates the overall normalized interface enrichment in the entire C-terminal half relative to the N-terminal half.

A bootstrapping strategy was used to calculate the error bars and *p*-values in Figure 2 and Figure S1. In short, the homomer or heteromer dataset used for each plot was randomly resampled 10^4 or 10^6 times, importantly allowing multiple instances of the same protein to be present in each interaction. Error bars were calculated as the standard deviation of the bootstrapping replicates, whereas *p*-values were calculated as the frequency in which the N- *vs.* C-terminal enrichments were greater than observed in the real dataset.

Note that there are two species which appear as outliers in the trend for C-terminal enrichment of interface residues. Only the hyperthermophilic anaerobic archaeon *Pyrococcus horikoshii* showed an opposite trend, which could reflect its unique habitat. In addition, although the enrichment appears to be much stronger in *Rattus norvegicus* than in other eukaryotes, it is likely that this is due to the types of rat proteins with structures available and the relatively small size of the dataset, as when the human orthologs of those rat proteins are considered, a similar enrichment is observed (Figure S1F).

Screening *Escherichia coli* homomers for their misassembly phenotypes. **Cell preparation**. The C-terminal GFP fusion version of the *E. coli* K-12 Open Reading Frame Archive library (ASKA) was grown in the original host strain *E. coli* K-12 AG1[51] in 96-well plates (growth conditions: 37°C, 280 rpm, LB medium). Following overnight growth, expression was induced for 2 hours by 0.1 mM IPTG in the fully-grown culture at 37°C. From the induced cultures 0.2 µL were carried over using a pin tool replicator into black CellCarrier-96 plates (PerkinElmer). In this plate each well had been supplemented with 100 µl of 1 µg/mL 4,6-diamidino-2-phenylindole (DAPI) in mineral salts minimal medium (MS-minimal) without any carbon source. Prior the microscopic analysis, cells were centrifuged down to the bottom of the 96 well plates.

Imaging. Microscopy was done using a PerkinElmer Operetta microscope. Four sites were acquired per well. Laser-based autofocus was performed at each imaging position. Images of two channels (DAPI and GFP) were collected using a 60x high-NA objective to visualize the cell and the aggregation states of the homomers, respectively. At every site and every fluorescent channel 5 images were taken at different z positions with 0.5 µm shifts. These images were used for a perfect focus algorithm. Cellular properties of about 1000 cells of each homomer-expressing strain were extracted from the images, including the localization of the GFP signal within the cell.

Image Analysis. Images were pre-processed using the CIDRE algorithm[5] to remove uneven illumination. A perfect focus algorithm was developed to locally select the best z image plane and create an image that contains the highest contrast cells. To identify cells and extract their properties, the CellProfiler

865 program[53] was used with custom modifications. First, image intensities were
866 rescaled. Then, cells were identified on the DAPI signal using Otsu adaptive
867 threshold and a Watershed algorithm to split touching cells. Cellular features
868 such as intensity, texture, and morphology were extracted.
869 **Phenotypic Classification using Machine Learning.** Supervised classification
870 of cells into predefined groups was done using the Advanced Cell Classifier
871 software (4). The cellular phenotypes were (*i*) no GFP signal (fluorescence level
872 equaled to that of the negative control without GFP) (*ii*) homogenous GFP signal
873 (cells show equally distributed GFP signal throughout the whole cell). Cells that
874 did not fit into these two categories were discarded. For the automated decision,
875 an artificial neural network method was used based on the Weka package[54].
876 Based on this cell classification, the homomers were assigned to one of the two
877 classes, depending on which phenotype was predominant in the cell population.
878 We considered a homomer as 'Dark' only if more than 50% of the cells were
879 dark. Where more than 50% of the cells showed homogeneous green
880 fluorescence, the homomer was classified as 'Green', which we refer to as
881 soluble homomers.
882 Due to the large number of cells, the classification of the phenotypes and image
883 analysis of the data was parallelized using high performance desktop PCs.
884 **Western Blot (WB).** Using western blot analysis, we tested if homomers with
885 predominantly 'Dark' cells are expressed in the samples and the lack of
886 fluorescence signal is not the result of compromised expression from the ASKA
887 plasmids. To this aim, the ASKA clones of all the 'Dark' homomers were
888 inoculated for overnight growth, and expression was induced for 2 hours in the
889 same way as for the image analysis described above. Following expression, cells
890 were harvested by centrifugation (~13,000g) and the pellets were re-suspended
891 in 2xSDS-sample buffer, adjusting its volume to the cell number. After boiling the
892 samples for 5 minutes, equal amount of total proteins in 5 µL were separated on
893 10% SDS-polyacrylamide gel (PAGE). Gels were either stained with Coomassie
894 Brilliant Blue (CBB) for justifying equal loading or transferred onto PVDF
895 membranes (Amersham, GE Healthcare Lifescience) proceeding further for
896 western blotting. Next, membranes were blocked in 5% milk powder-0.05%
897 Tween20 in TBS (25mM Tris-Cl, pH 8.0, 150 mM NaCl) buffer (TBST) for an hour

at room temperature (RT). Anti-GFP (Chromotek) was used as primary antibody diluted in 5% milk powder-TBST (1:1000) buffer for overnight incubation at 4^0C. After washing with TBST buffer to remove the excess of antibody, membranes were incubated with secondary antibody diluted in 2.5 % milk-powder-TBST buffer (1:10000) for an hour on RT. After washing the membranes in TBST buffer, signals were developed by a standard chemiluminescent western blot detection method (Thermo Scientific). Signals were converted into black and white images and then the Image J program was used for quantifying the western blot results (degradation products were not counted). Band area was then corrected by eliminating the background value and was normalized to a relative value with the positive control present in each western blots (Supplementary Data Set 1). To determine the expression level threshold below which GFP fluorescence cannot be detected, we also performed Western blot assays on a set of 23 'Green' homomers. 'Dark' homomers below this expression level (*i.e.* with weak or no expression) were removed from the dataset and only those 'Dark' homomers were considered as co-translationally aggregating in the downstream analysis, which gave a relative band area of ≥0.25 in the western blot analysis. GFP positive degradation products were not counted. Information on all 'Green' and 'Dark' homomers can be found in Supplementary Data Set 1.

Cell culture and expression. A single colony was picked into a 2xTY media with 40mg/l kanamycin and allowed to grow overnight (O/N) at 37°C. A fresh media was inoculated with the O/N culture and let grow while vigorously shaken until it reached OD_{600}=0.4-0.6. The culture was then induced with a final concentration of 0.1mM Isopropyl β-D-1-thiogalactopyranoside (IPTG). Cells were left to grow at 37°C or 18°C for 3hr or 15hr, respectively (for flow cytometery experiments), or for 4-6hr at 37°C for protein purification. For co-expression experiments, the media had both kanamycin and ampicillin for a positive selective of cells containing the plasmid with the construct (Tet-SL-YFP or Mono-SL-YFP) and plasmid containing the TetD peptide, respectively.

Protein purification. As described previously[55], the induced cells were harvested and left at -20°C. The frozen cells were then resuspended in a cold ice lysis buffer [50mM NaCl, Tris pH=7.2, ®-mercaptoethanol (*Sigma-Aldrich*), Complete Protease Inhibitor (*Roche*), and RNAse and DNAse from Bovine pancreas (*Sigma-Aldrich*)] and sonicated. After centrifugation, the supernatant was loaded on a 1ml Anionic column (*GE Healthcare*) or 5ml HisTrap column (*GE Healthcare*) and eluted with a gradient of 1M NaCl or 500mM Imidazole buffer, respectively. For the tetrameric constructs the proteins were eluted at high Imidazole concentration (>150mM). The elution was dialyzed against 150mM NaCl, 20mM Tris pH=7.2 buffer and loaded on Gel Filtration HiLoad 16/600 Superdex 200 (*GE Healthcare)* connected to ÄKTAPurifier FPLC systems (*GE Healthcare*). All constructs were 90-95% pure as determined by 4-12% Bis-Tris SDS page gel.

Western Blot (WB). Cells were grown and expressed as described above. The cells were shortly centrifuged (~13,000g), frozen at -20°C and resuspended on ice using Tris-buffer. The cells were centrifuged using a temperature controlled bench centrifuge (*Eppendorf*) at 4°C for 30min. Pellet and supernatant were separated. A second round of resuspension was conducted followed by centrifugation to verify that all soluble proteins were extracted. No additional protein was found at that stage. Samples were heated at 95°C for 5min with loading buffer (*NuPAGE Novex, LifeTechnologies*). The exact same volume was loaded into 4-12% Bis-Tris SDS page gel (*NuPAGE Novex, invitrogen*) in MES buffer (2-(N-morpholino)ethanesulfonic acid). Each sample was run twice on different gels to eliminate the possibility of loading inconsistency and other technical issues. Blotting and transfer was conducted with iBlot® Gel Transfer (*Life Technologies*). The membranes were incubated in PBST [PBS, 0.1% Tween (v/v)] and 2% BSA for blocking O/N or 2hr at 4°C or at room temperature, respectively. Primary rabbit monoclonal antibody (Anti-HA Tag, *Millipore*) was diluted in PBST (1:5,000) and was detected using a secondary antibody that was diluted in PBST (1:1000). To remove the excess of secondary antibodies, we washed with PBST. For detection we used Amersham ECL Western Blotting detection kit (*GE Healthcare, Life Sciences*) and V3 Western Workflow (*GE*

962 *Healthcare, Life Sciences*) in 1-10sec increments. Measurements were repeated
963 at least three times showing consistency between repeats.
964
965 **Flow Cytometry.** The cultures were grown and induced as described above. The
966 overnight expression of each construct was divided into three replicates before
967 induction, and each sample was measured separately. Each measurement was
968 triplicated and repeated at least three times on different days using different
969 colonies. The cultures were incubated at 18°C for 10-15hr before being
970 measured. The culture was centrifuged briefly (~15sec at 13,000g),
971 resuspended and diluted with PBS thereafter. The sample was measured using
972 BD LSRII and a BD LSRFortessa (*Becton Dickinson*) with a 488nm laser and
973 detection at 525nm. Samples were sent for sequencing for verification after
974 measurements took place.
975
976 **Flow cytometry data analysis.** Data was analysed using FlowJo software
977 (version 10.0.6). To discriminate doublets, SSC-H was aligned against SSC-W,
978 and the appropriate gating was applied. Then the population was divided into
979 fluorescent and non-fluorescent sub-populations, where the median value of the
980 former was extracted. Fluorescence levels of the same variants measured on the
981 same day were averaged, and the ratio between the tetrameric and monomeric
982 pair, i.e. of the same sub-library, was calculated. Measurements were repeated
983 using different colonies on different days. The tetrameric-to-monomeric values
984 were then averaged. Standard deviation (Excel) was calculated and t-test used to
985 determine significance.
986
987 **Confocal Microscopy**. Cells were grown as described above and were induced
988 for 4-7hr, then washed with PBS three times and allowed to adhere in pre-
989 treated slides for a few minutes, images were acquired soon after. The images
990 were taken using Zeiss710 (Carl Zeiss) with an objective of 63x, an excitation
991 laser of 525 nm and emission window between 581nm and 750nm. At least 100
992 cells were captured for each strain and growth condition. We used a Leica DMRB
993 microscope equipped with a Leica DC-200 camera. Images were taken at a

magnification of 160x using a filter for GFP excitation (450–495nm) and an emission filter (515–560nm). Samples were sent to sequencing for strain verification after measurements took place.

Native Mass Spectrometry. Intact mass spectrometry measurements were performed on a Waters Synapt (first generation) HDMS system modified for high mass transmission as previously described[56]. Samples were buffer exchanged into 200 mM ammonium acetate solution using Bio-spin 6 (Bio-Rad) columns. Typically, 3 μL of sample was loaded into gold-coated capillaries prepared in-house[57] and mounted into a static nanospray source allowing the application of high voltage to the capillary. The instrument operating parameters were: capillary voltage 1.1-1.5 kV, sample cone 60-100 V, extraction cone 3 V, trap/transfer collision cells were maintained at $5.52x10^{-2}$ mbar and 10-20 V collision voltage, Backing pressure $6x10^{-3}$ mbar. Data was analysed using the MassLynx software. Tandem MS experiments were performed applying collision-induced dissociation from the trap collision cell of the instrument on the parent ions isolated using the quadrupole.

Tecan200. The OD_{600} of the overnight cultures was measured and diluted accordingly to a final value of 0.1 in a fresh 2xTY solution with the suitable antibiotics. To each strain IPTG was added to a final concentration of 0.1mM. The samples were then allocated in triplicates in 96-wells plate and measured while growing for 24hr. The measurements were repeated on different days using different colonies. Using i-control™ (*Tecan*) temperature (37°C) and the shaking-reading cycles were determined. Absorbance measurements were at 600nm and fluorescence at 485±9nm/535±20nm wavelength for emission/excitation, respectively.

Tecan analysis (R+). The triplicate average of OD_{600} values was taken at each time point, and growth curves were fitted with a spline algorithm from the 'grofit' package in R. Confidence intervals were computed by bootstrapping.

***In vivo* Luciferase expression.** The cultures were grown and induced as

described above. About 10uL of the induced media was lysed in 50μL of Passive Lysis Buffer (*Promega*). Expression levels of active luciferase were evaluated by luminescence signal of 2μL aliquot of the lysate. Dual-Glo Luciferase Substrate (*Promega*) was dispensed to the lysate and the luminescence signal was measured using a microwell plate reader (Varioskan Flash, *Thermo Scientific*). Total protein level was evaluated using Western Blot analyses using anti-HA tag antibody and DyLight 649-conjugated ant-mouse IgG as the first and second antibodies, respectively. Blotting and binding reaction of antibodies were performed using iBlot Gel Transfer and iBind Western System (*Life Technologies*). Fluorescence signal on the membrane was imaged by fluorescence scanner (FLA-5100, *GE Helthcare*) using 633 nm excitation and 665 nm emission, and signal intensities were calculated by the Multi Gauge software (*Fuji Film*).

PURE System. DNA templates encoding the reporter gene constructs were prepared by PCR amplification from appropriate plasmids using T7 promoter and T7 terminator primers. mRNAs were transcribed *in vitro* using Thermo T7 Transcription Kit (*Toyobo*) and purified using RNeasy Mini Kit (*Qiagen*) followed by Centrisep Spin Column (*Princeton*). PURE system solution (*PURExfrex*), in which Release Factor 1, Release Factor 2, and ribosomes were removed, was purchased from GeneFrontier Corp. For a polysomic translation, mRNA and ribosome (*GeneFrontier*) were mixed in the reaction solution of 60nM and 3.0μM, respectively. For a monosomic translation, mRNA and ribosome (*GeneFrontier*) were mixed in the reaction solution of 1.8μM and 0.6μM, respectively. It should be noted that when constructs with YFP were translated, tRNA concentration in the reaction mixture was reduced to 20% of that specified by the manufacturer. Translation reaction was performed at 37°C for 15min, and terminated by addition of 20μM puromycin followed by 10min incubation at 37°C. To evaluate expression levels of active florescent proteins, aliquots (2μL) of translated products were diluted in PBS containing 0.01% Tween 20 (50μL) and fluorescence signals of YFP and GFP (fGFP and GFP) were measured immediately or after 24hr incubation at 25°C, respectively, using

488nm excitation and 530nm emission. To evaluate expression levels of active luciferase, aliquots (2μL) of translated products were diluted in PBS containing 0.01% Tween 20 (70μL), and luminescence signals were measured after addition of Dual-Glo Luciferase Substrate. Total protein expression levels in the translated products were quantified by Western Blotting according to the above procedure. Correct folding of each construct was calculated as the ratio between the fluorescence signals and total protein as established by WB. The value of the tetrameric construct was then divided by the monomeric constructs of the same reporter gene. All measurements were repeated at least three times on different days, and the results presented are the average of those repeats. It should be noted that when luciferase was used as a reporter-gene, western blotting was performed using Alkaline Phosphatase-conjugated anti-mouse IgG as the second antibody, and colorimetric signal was obtained by using Western Blue Stabilized Substrate for Alkaline Phosphatase (*Promega*).

PURE-Chaperones. Similarly to the above protocol, to the PURE*frex* mix (polysomic conditions) we added DnaK/DnaJ/GrpE mixture (*GeneFrontier*) or GroEL/GroES (*GeneFrontier*), or trigger factor (a generous gift of Dr. Guenter Kramer) as previously reported[58], to evaluate the effect of these chaperones on the folding of reporter proteins. Signal as fluorescence or luminescence, and WB for each measurement were measured for samples with or without the different chaperone mix hues / TF. Each experiment was repeated three times, the averaged values of the different experiments are presented in Figure S8. The ratio of with/without chaperones result are presented in Figure 5D.

Simulation. Molecular simulations of the conformational behavior of the nascent-chain constructs were performed using protocols similar to those used in previous work by us[59]. Both the nascent-chains and the ribosomes were modeled using residue-level, coarse-grained representations and the conformational dynamics of both molecules were simulated using the technique of Brownian dynamics[60].

Structures used in the simulations: 70S ribosomes were modeled using the *E. coli* structure solved by Agirrezabala *et al.*[61]. Two such ribosomes were arranged

in the "i:i+3" geometry identified in tomographic reconstructions of *E. coli* polyribosomes[62]; this arrangement was selected for simulation as it places the exit tunnels for the two nascent-chains in closest proximity and thereby maximizes the chances of observing co-translational assembly events. The structures of all p53/YFP constructs were built using the PDB files 1GFL for YFP[63] and 1C26 for the p53 oligomerization domain[64]. Homology modeling was performed using the program SwissModel[65] and missing loops were constructed using the program[66]. The following constructs were simulated: Tet-SL-YFP, YFP-SL-Tet, and Tet-LL-YFP; structures of these three constructs in their modeled, native conformations are shown in Figure S9. Note that due to the computational expense of the simulations we considered only the formation of dimeric constructs in the simulations.

Energetic calculations for the simulations: As in our previous work[59], all nascent-chain constructs were modeled using standard molecular mechanics bond stretch, angle and dihedral terms with steric interactions applied to prevent pseudoatoms from overlapping with each other. To enable the native constructs to fold correctly, additional favorable Lennard-Jones potential functions were used to reward the formation of known native contacts. As in our previous work[59] native contacts were defined as any pair of residues for which any pair of heavy atoms were within 5.5 Å in the native state. For the YFP domain, the energy well-depth assigned to native contacts was set to 0.6 kcal/mol, a value that we have previously shown provides a good description of the (intra-molecular) folding thermodynamics of typical single domain proteins[59]. For the p53 oligomerization domain, a somewhat deeper well-depth of 1.2 kcal/mol was used to ensure that intermolecular contacts, if formed, remained stable during the simulations; these native intermolecular contacts were defined using chains A and C of the crystal structure of the p53 oligomerization domain[67]. Electrostatic interactions between the nascent-chains and with the ribosomes were modeled using the Debye-Hückel approximation, with a cutoff of 25 Å and an assumed ionic strength of 150 mM; charges were assigned to each residue using the Henderson-Hasselbalch equation assuming a pH of 7.6 and model pKa values taken from a literature survey of pKa values in proteins[68].

All simulations were performed using software written in-house and using the Langevin dynamics algorithm developed as an extension of the Ermak-McCammon algorithm by the Geyer group[69]. Simulations were performed at 310K, with the solvent dielectric constant and viscosity set to the corresponding experimental values for water. All pseudoatoms of the nascent-chain constructs and the ribosomes were allowed to move in the simulations, but harmonic restraints were applied to the ribosome atoms to ensure that its overall structure was maintained. During the periods in which synthesis of the nascent-chains was simulated, the C-terminal four residues of each nascent-chain were harmonically restrained to modeled positions in the ribosome peptidyltransferase active site. To ensure rapid conformational diffusion of the unrestrained parts of the nascent-chains, and of the entire chain at the completion of synthesis, their intramolecular hydrodynamic interactions were explicitly modeled in the simulations using the Rotne-Prager-Yamakawa level of theory as implemented in our previous work[70]. The diffusion tensors describing these hydrodynamic interactions were updated every 100 ps and the Cholesky decompositions required for the generation of correlated random displacements were calculated using the fast parallelized code developed by Hogg *et. al*. A time-step of 50 fs was used in all simulations, with new amino acids added to the growing nascent-chains every 160,000-simulation steps, *i.e.*, every 8 ns. This is clearly much faster than occurs in real life, but this issue is mitigated by the fact that the simulated folding timescales of the domains are also similarly accelerated relative to their experimental values[59].

We have estimated previously that in the "i:i+3" arrangement of the two ribosomes, the nascent-chain at ribosome "i" is likely to be ~72 amino acids longer than the chain at ribosome "i+3"[62]; we therefore allowed synthesis of the first nascent-chain to reach 72 amino acids before synthesis of the second chain was begun. Once fully synthesized, both nascent-chains were free to leave the ribosome. During each simulation, the numbers of intramolecular and intermolecular native contacts within and between the nascent-chains were monitored in order to determine the extents of folding and assembly; contacts were considered to have been successfully formed if two residues were within a

factor of 1.2 of their separation distance in the native structures. And misassembly was defined as a stable non-native contact between two YFP domains. To obtain an estimate of the uncertainties in the simulated behaviors, all constructs were simulated 20 times, with a different series of random displacements[60] ensuring independence of the trajectories.

Structural analysis of *E. coli* multidomain homomers. All *E. coli* protein structures and their corresponding interface residues were identified as above. Protein residue positions were mapped to their respective UniProt protein position and their protein domain definitions according to the Structural Classification Of proteins (SCOP) and the Protein Family Database (Pfam) using the Structure Integration with Function, Taxonomy and Sequences resource (SIFTS)[71] API in a customized Python script. All protein complex structures that have more than one SCOP or PFAM domain and are homomers were extracted (196 *E. coli* multi-domain protein structures, out of which 150 were homomers). To only identify proteins where the whole protein rather than single domains or fragments have been crystallized, only structures that cover at least 95% of the UNIPROT sequence were used (91 protein structures).

To determine which of these proteins have their interface localized in an 'oligomerization domain' and resemble the architecture of the p53-GFP construct in this study, for every position of the protein interface structures, the relative interface contribution of each domain (defined as the fraction of total buried surface area (BSA) provided by all residues of a domain) was computed. 5 structures had >95% of their interface region in one domain (as defined by Pfam and SCOP). The linker-length between the domains was determined as the uniprot residues that separate the respective Pfam domains of each protein structure.

The isolated structures were mapped to their translational folding-rates using the data generated by O'Brien[72] *et. al.* by using the generated PDB to uniprot mapping. Four of the five proteins had translational folding-rates associated.

Protein complex immunoprecipitation (Co-IP). A strain with an 'empty

vector' and the strains that express the tetrameric N-terminus (Tet-SL-YFP) and C-terminus (YFP-SL-Tet) constructs were harvested a few hours after induction. Then the cells content were mixed with magnetic beads covered with anti-HA antibodies. Both constructs had a C-terminal HA-tag (Pierce HA-Tag Magnetic IP/Co-IP Kit). Proteins were eluted and run on a SDS gel. Bands that appeared to be different between the samples were extracted (1-2mm) and the excised protein gel pieces were placed in a well of a 96-well microtitre plate and destained with 50% v/v acetonitrile and 50mM ammonium bicarbonate, reduced with 10mM DTT, and alkylated with 55mM iodoacetamide. After alkylation, proteins were digested with 6ng/μL Trypsin (*Promega*) overnight at 37°C. The resulting peptides were extracted in 2% v/v formic acid, 2% v/v acetonitrile. The digest was analyzed by nano-scale capillary LC-MS/MS using an Ultimate U3000 HPLC (*Thermo Scientific Dionex*) to deliver a flow of approximately 300nL/min. A C18 Acclaim PepMap100 5μm, 100 μm x 20 mm nanoViper (*Thermo Scientific*), trapped the peptides prior to separation on a C18 Acclaim PepMap100 3μm, 75μm x 150mm nanoViper (*Thermo Scientific Dionex*). Peptides were eluted with a gradient of acetonitrile. The analytical column outlet was directly interfaced via a modified nano-flow electrospray ionisation source, with a hybrid dual pressure linear ion trap mass spectrometer (*Orbitrap Velos*, *Thermo Scientific*). Data dependent analysis was carried out, using a resolution of 30,000 for the full MS spectrum, followed by ten MS/MS spectra in the linear ion trap. MS spectra were collected over an m/z range of 300–2000. MS/MS scans were collected using threshold energy of 35 for collision induced-dissociation. LC-MS/MS data were then searched against a protein database (*UniProtKB*) using the Mascot search engine software (*Matrix Science*). Database search parameters were set with a precursor tolerance of 5 ppm and a fragment ion mass tolerance of 0.8 Da. Two missed enzyme cleavages were allowed and variable modifications for oxidized methionine, carbamidomethyl cysteine, pyroglutamic acid, phosphorylated serine, threonine and tyrosine were included. MS/MS data were validated using the Scaffold software (*Proteome Software Inc.*). All data were additionally interrogated manually.

The influence of chaperons on homomeric and heteromeric complexes in *E. coli* was investigated using the dataset from Ref[72]. The depletion of misfolded

homomeric and heteromeric protein complexes from the soluble fraction of *E. Coli* mutants with ΔKJT deletion (DnaK/DnaJ and TF are deleted) was visualized using R scripts (data from Table S8 in "Change in abundance in insoluble fraction"[73]). In addition, the interaction of homomeric and heteromeric complex proteins with DnaK (PD/BG ratio, data from Table S2 in same paper[73]) was analyzed. The relative frequencies were normalized to account for the number of homomeric and heteromeric complexes.

Methods Only References

48. Levy, E.D. PiQSi: protein quaternary structure investigation. *Structure* **15**, 1364-7 (2007).
49. Marsh, J.A. & Teichmann, S.A. Protein flexibility facilitates quaternary structure assembly and evolution. *PLoS Biol* **12**, e1001870 (2014).
50. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195-201 (2006).
51. Kitagawa, M. et al. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res* **12**, 291-9 (2005).
52. Smith, K. et al. CIDRE: an illumination-correction method for optical microscopy. *Nat Methods* **12**, 404-6 (2015).
53. Carpenter, A.E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* **7**, R100 (2006).
54. Hall, M. et al. The WEKA data mining software: an update. SIGKDD Explor Newsl 11 (1): 10–18. doi: 10.1145/1656274.1656278. (2009).
55. Natan, E. & Joerger, A.C. Structure and kinetic stability of the p63 tetramerization domain. *J Mol Biol* **415**, 503-13 (2012).
56. Sobott, F., Hernandez, H., McCammon, M.G., Tito, M.A. & Robinson, C.V. A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. *Anal Chem* **74**, 1402-7 (2002).
57. Hernandez, H. & Robinson, C.V. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat Protoc* **2**, 715-26 (2007).
58. Niwa, T. et al. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc Natl Acad Sci U S A* **106**, 4201-6 (2009).
59. Elcock, A.H. Molecular simulations of cotranslational protein folding: fragment stabilities, folding cooperativity, and trapping in the ribosome. *PLoS Comput Biol* **2**, e98 (2006).
60. Ermak, D.L. & McCammon, J. Brownian dynamics with hydrodynamic interactions. *J Chem Phys* **69**, 1352-1360 (1978).
61. Agirrezabala, X. et al. Structural insights into cognate versus near-cognate discrimination during decoding. *EMBO J* **30**, 1497-507 (2011).

62. Brandt, F. et al. The native 3D organization of bacterial polysomes. *Cell* **136**, 261-71 (2009).
63. Yang, F., Moss, L.G. & Phillips, G.N., Jr. The molecular structure of green fluorescent protein. *Nat Biotechnol* **14**, 1246-51 (1996).
64. Jefferys, B.R., Kelley, L.A. & Sternberg, M.J. Protein folding requires crowd control in a simulated cell. *J Mol Biol* **397**, 1329-38 (2010).
65. Marsh, J.A. et al. Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* **153**, 461-70 (2013).
66. Xiang, Z., Soto, C.S. & Honig, B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci U S A* **99**, 7432-7 (2002).
67. Jeffrey, P.D., Gorina, S. & Pavletich, N.P. Crystal structure of the tetramerization domain of the p53 tumor suppressor at 1.7 angstroms. *Science* **267**, 1498-502 (1995).
68. Antosiewicz, J., McCammon, J.A. & Gilson, M.K. The determinants of pKas in proteins. *Biochemistry* **35**, 7819-33 (1996).
69. Winter, U. & Geyer, T. Coarse grained simulations of a small peptide: Effects of finite damping and hydrodynamic interactions. *Journal of Chemical Physics* **131**(2009).
70. Frembgen-Kesner, T. & Elcock, A.H. Striking effects of hydrodynamic interactions on the simulated diffusion and folding of proteins. *Journal of chemical theory and computation* **5**, 242-256 (2009).
71. Velankar, S. et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* **41**, D483-9 (2013).
72. O'Brien, E.P., Vendruscolo, M. & Dobson, C.M. Prediction of variable translation rate effects on cotranslational protein folding. *Nat Commun* **3**, 868 (2012).
73. Calloni, G. et al. DnaK functions as a central hub in the E. coli chaperone network. *Cell Rep* **1**, 251-64 (2012).

Acknowledgment

We are grateful to Günter Kramer and Bernd Bukau for their generous gift of Trigger Factor protein, and A. Drummond for the generous gift of plasmids. We would also like to thank L. Byung-Gil for useful advice and to N. Sanchez De Groot for technical support. We thank C Vogel, MT Burgas and E Arbely for helpful suggestion and critical reading. EN would like to thank Nina Weiner and the ISEF foundation for their support. MMB, TF and GC are supported by the Medical Research Council (MC_U105185859). TF was also supported by the Boehringer Ingelheim Fond. BP and CP would like to thank 'Lendület' Programme of the Hungarian Academy of Sciences and the Wellcome Trust for supporting this work, and the European Research Council (CP). BK is supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and NKFI 120220. PH would like to thank the National Brain Research Programme and the TEKES Finland Distinguished Professor Grant for their support. SAT thanks the Lister Institute, the MRC, the EMBL-European Bioinformatics Institute and the Wellcome Trust Sanger Institute. NS and TE were partly supported by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), mostly Innovative Areas of "Chemistry for Multimolecular Crowding in Biosystems" (JSPS KAKENHI Grant No. JP17H06351), and MEXT-Supported Program for the Strategic Research Foundation at Private Universities (2014-2019) and The Hirao Taro Foundation of KONAN GAKUEN for Academic Research. JM is supported by an MRC Career Development Award (MR/M02122X/1). CR is supported by the Medical Research Council, Grant Reference MR/N020413/1. LHV was supported by EMBO (award number ALTF 698-2012), Directorate-General for Research and Innovation (FP7-PEOPLE-2010-IEF, ThPLAST 274192) and an EMBL Interdisciplinary Postdoctoral fellowship, supported by H2020 Marie Skłodowska Curie Actions. BP and HP acknowledge funding from GINOP-2.3.2-15-2016-00026. AHE's work was supported by the National Institutes of Health through grant R01 GM099865. This work is dedicated to Jakob Natan and Shalom Marciano.

Contributions

The study was conceived by EN and SAT
The study was coordinated by EN and SAT.
The experiments were designed by EN, LHV, BK, BP, CP and PH.
The experiments were conducted by EN, TE, NS, AHE, BK, LD, EŐ and ZM.
Bioinformatic analysis was conducted by TF and JAM.
Simulations were run by AHE.
Machine learning analysis was conducted by PH.
Data analysis was conducted by EN, TE, AHE, TF, BK, GF, HP, BP CP and GC.
The manuscript was written by EN and SAT with contributions from all authors.

Conflict of interest

The authors declare that they have no competing financial interests.

Data and Code availability

1. Code and datasets for "*Screening Escherichia coli homomers for their misassembly phenotypes*" section is available at doi: https://bitbucket.org/feketegergo/n-terminal-enrichment-analysis-of-e-coli-homomers/src
2. Code and datasets for "*Structural analysis of interface location*" section is available at doi: http://hdl.handle.net/10283/2918
3. Code use to generate video S1-3 and Figure S9 is available upon request from the authors (Adrian H Elcock).
4. Videos are available at doi: https://figshare.com/s/48830eb9d72a80065d4a

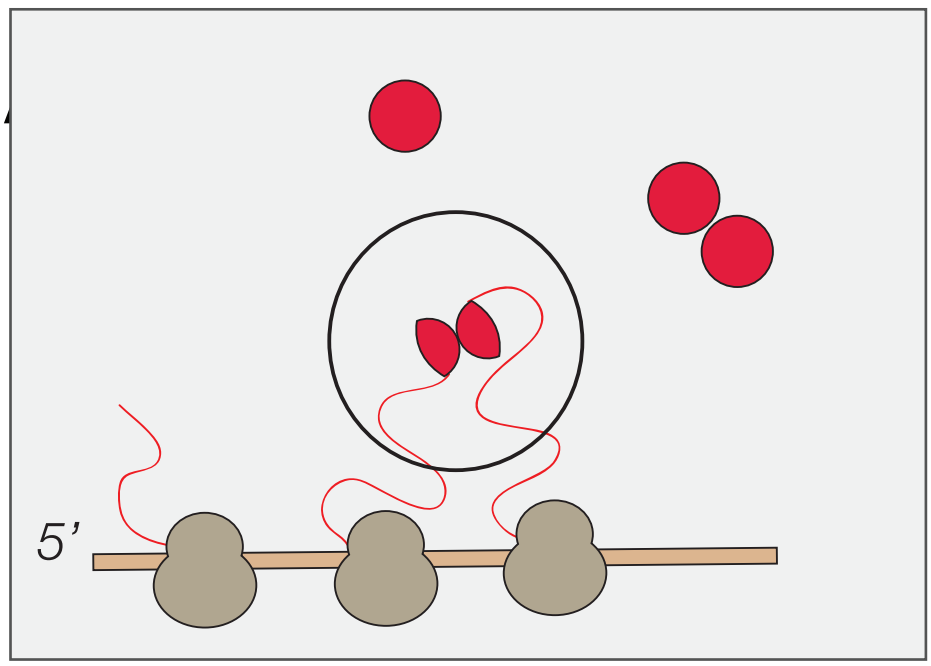
The data that support the findings of this study are available from the corresponding author upon reasonable request.

A Life Sciences Reporting Summary for this article is available.

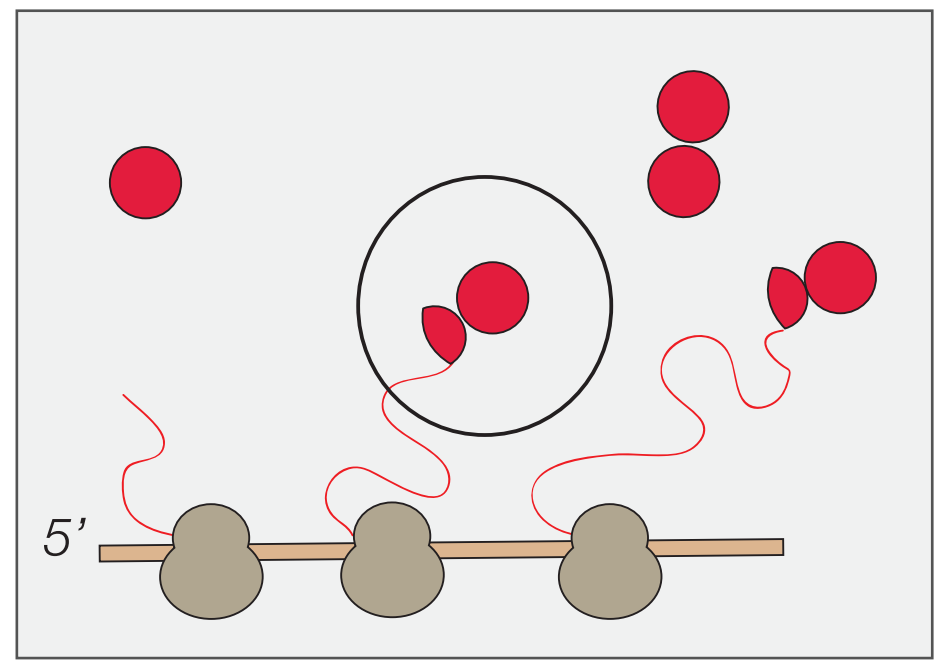
Supplementary Data Sets 1-4 for Figure 2C-E, 5A-C, and S10 are available with the paper online

1366
1367

Homo-Oligomerization of polypeptides from same mRNA

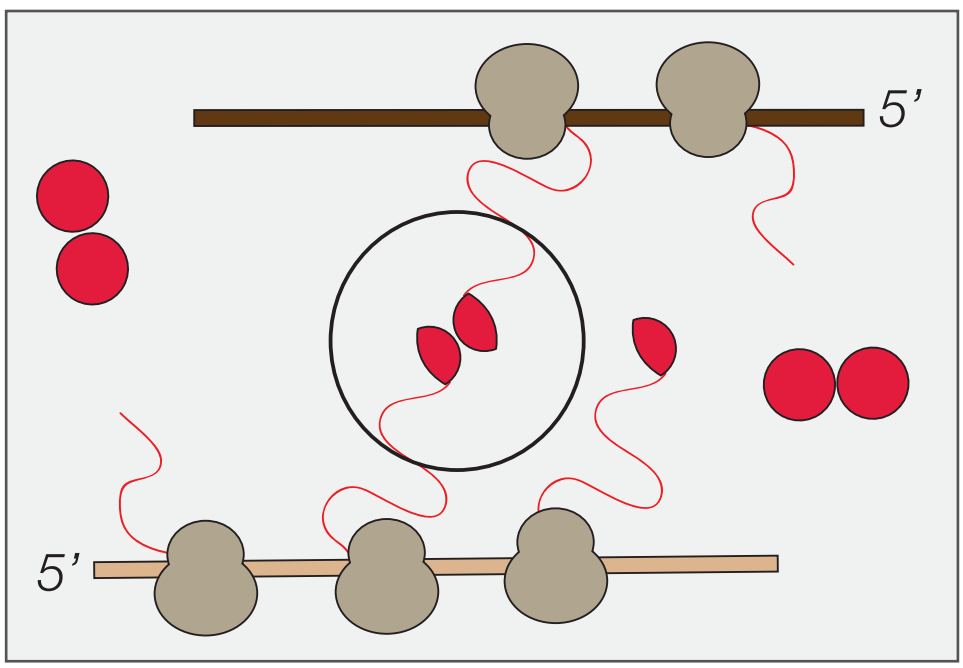


Same polyribosome oligomerization

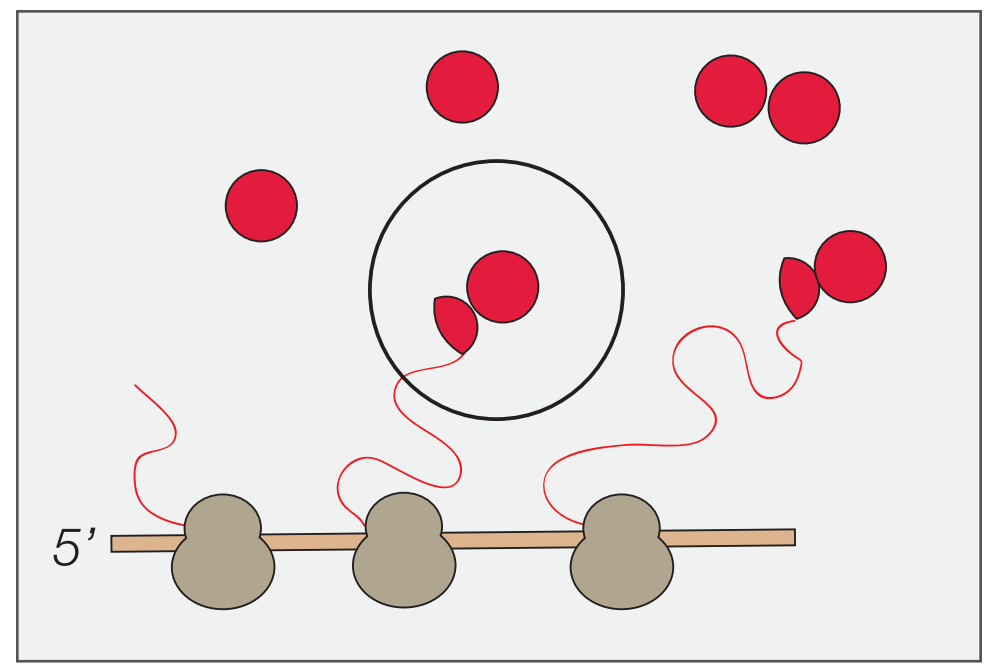


Oligomerization of a nascent chain and mature protein

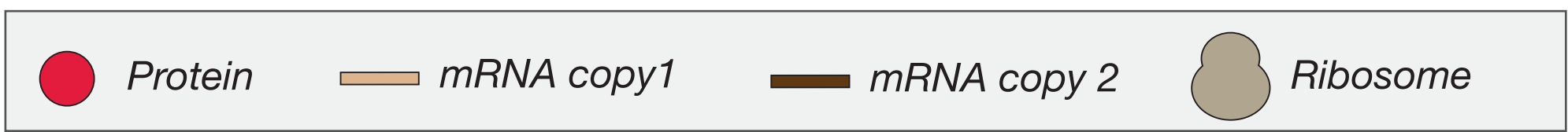
Homo-Oligomerization of polypeptides from different mRNA copies



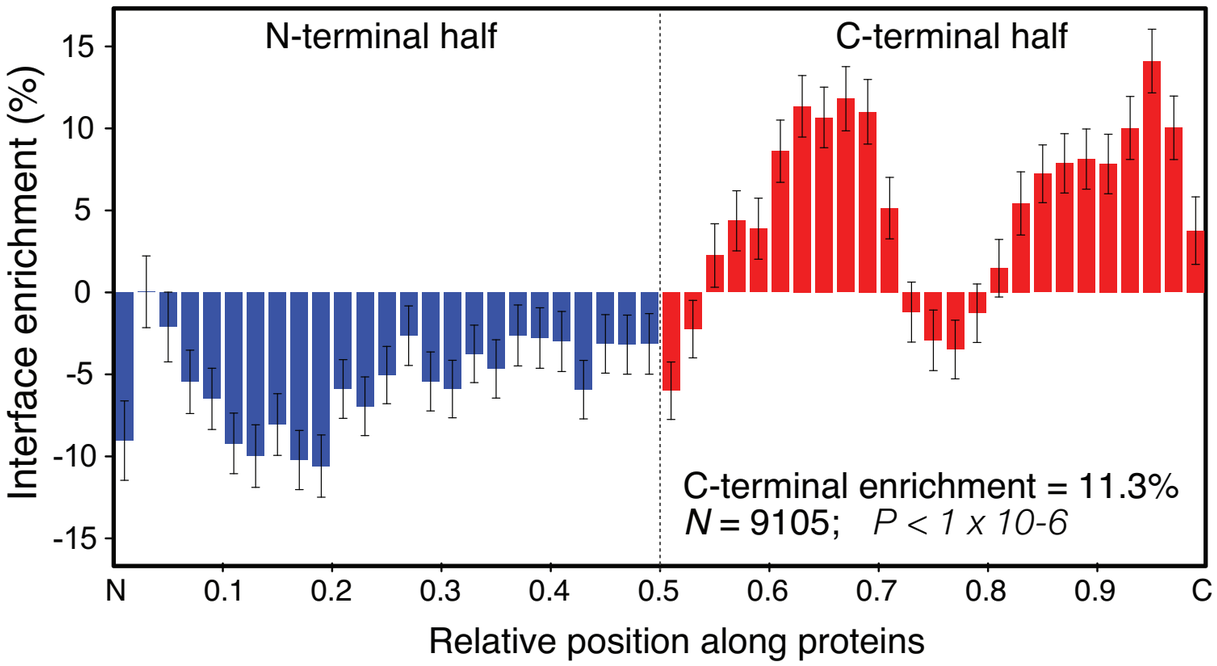
Different polyribosomes oligomerization



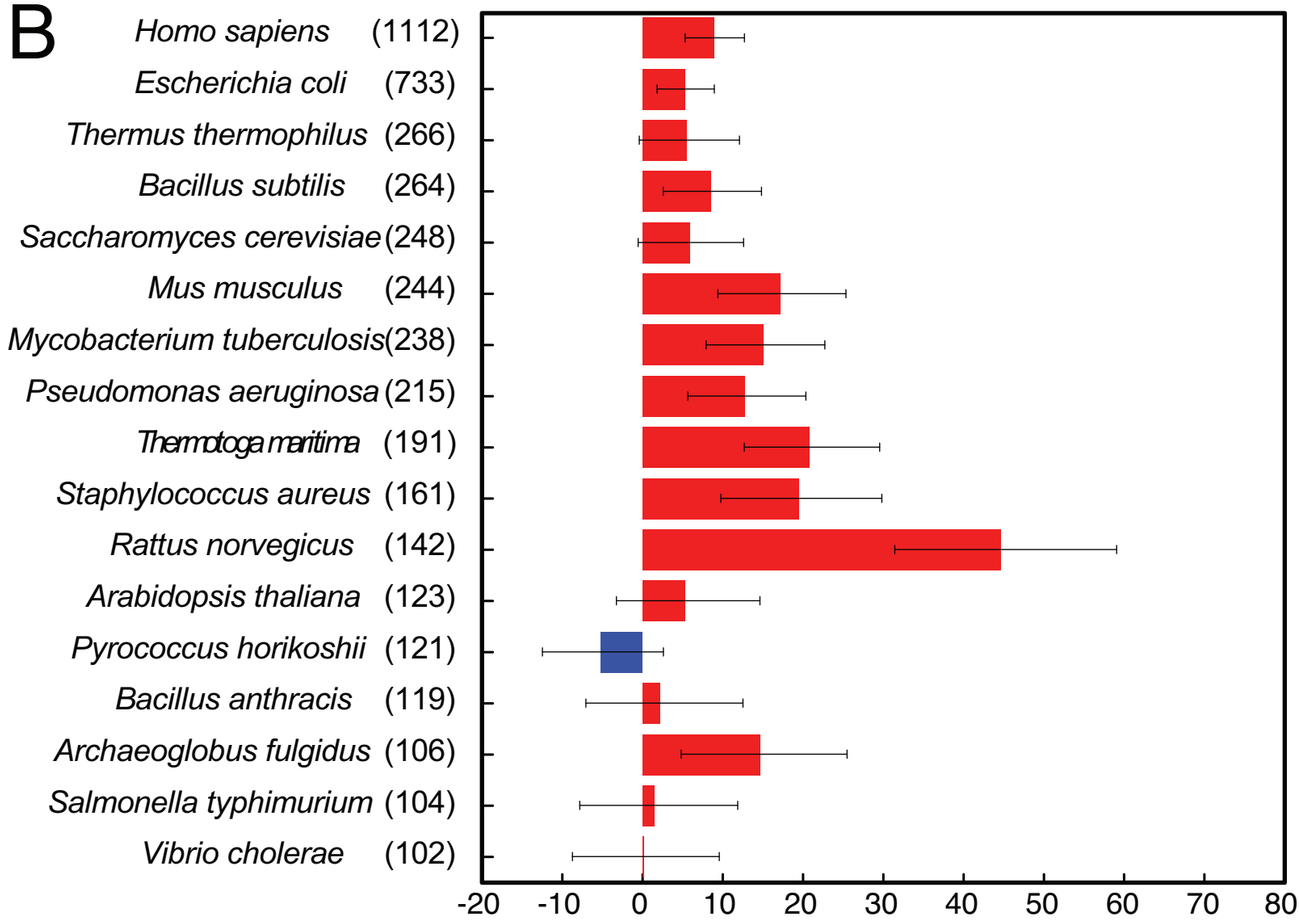
Oligomerization of mature protein and a nascent chain



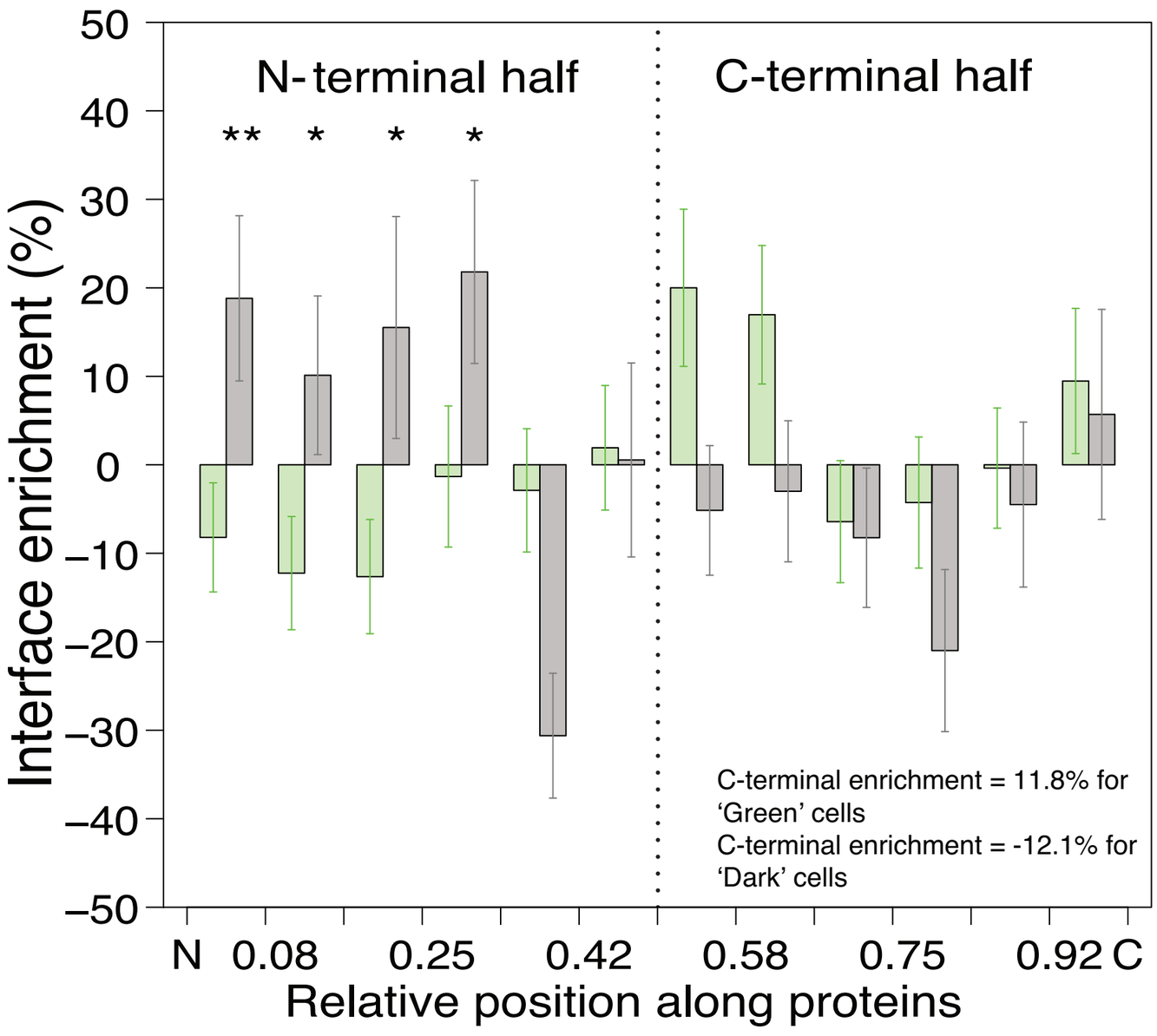
A


N-terminal half
C-terminal half
Interface enrichment (%)
15
10
5
0
-5
-10
-15
C-terminal enrichment = 11.3%
N = 9105; P < 1 x 10-6
N
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
C
Relative position along proteins

B

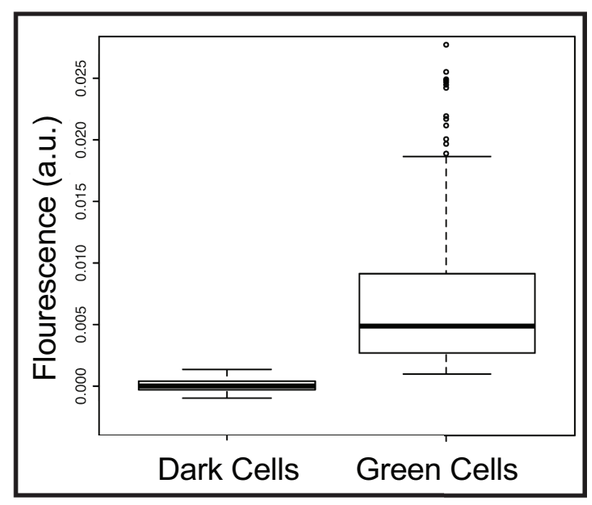

Homo sapiens (1112)
Escherichia coli (733)
Thermus thermophilus (266)
Bacillus subtilis (264)
Saccharomyces cerevisiae (248)
Mus musculus (244)
Mycobacterium tuberculosis (238)
Pseudomonas aeruginosa (215)
Thermotoga maritima (191)
Staphylococcus aureus (161)
Rattus norvegicus (142)
Arabidopsis thaliana (123)
Pyrococcus horikoshii (121)
Bacillus anthracis (119)
Archaeoglobus fulgidus (106)
Salmonella typhimurium (104)
Vibrio cholerae (102)
-20
-10
0
10
20
30
40
50
60
70
80

C

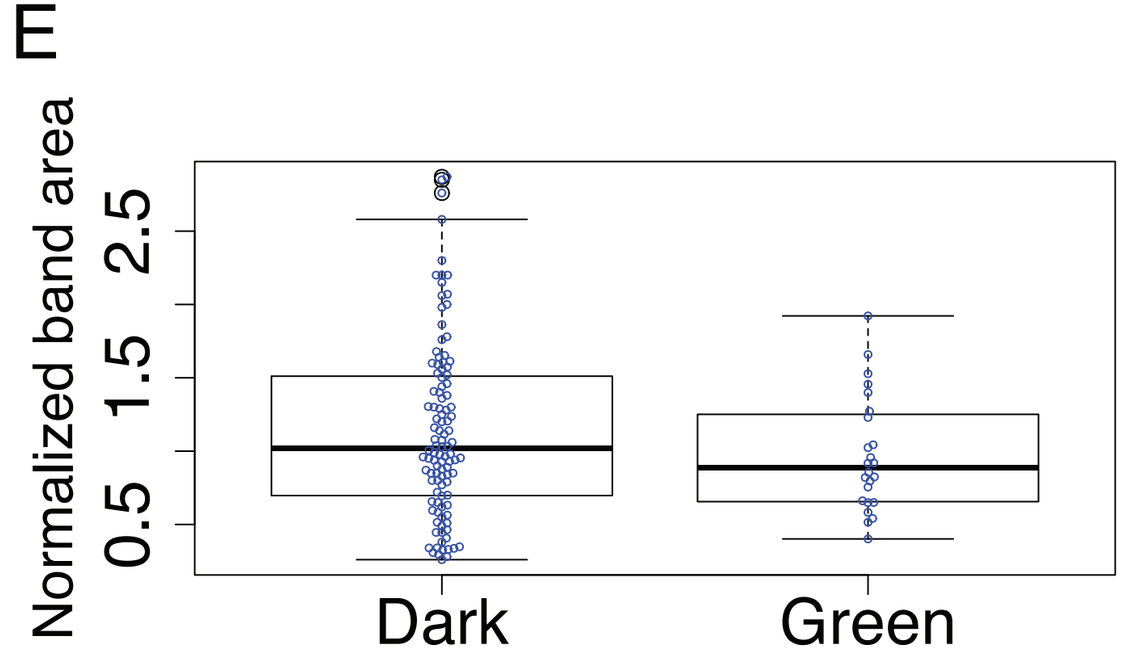

N-terminal half
C-terminal half
**
*
*
*
Interface enrichment (%)
50
40
30
20
10
0
–10
–20
–30
–40
–50
C-terminal enrichment = 11.8% for 'Green' cells
C-terminal enrichment = -12.1% for 'Dark' cells
N
0.08
0.25
0.42
0.58
0.75
0.92
C
Relative position along proteins

D

Dark cells
Green cells


Flourescence (a.u.)
0.000
0.005
0.010
0.015
0.020
0.025
Dark Cells
Green Cells

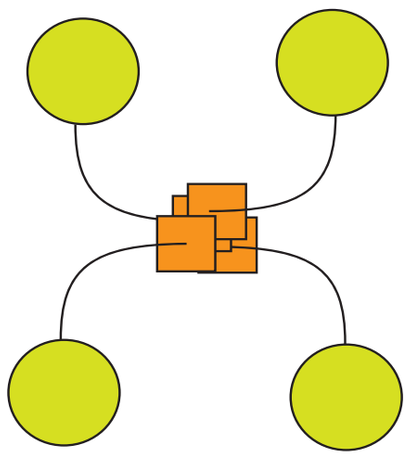
E

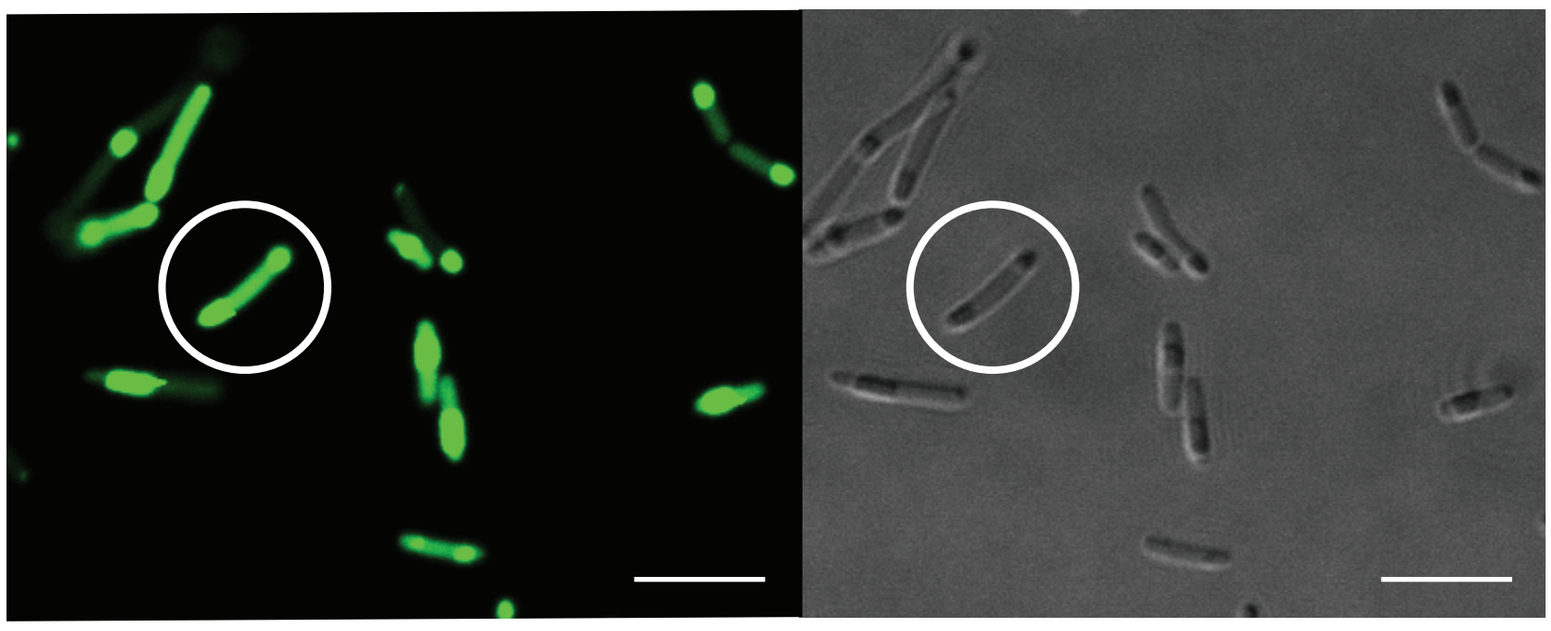

Normalized band area
0.5
1.5
2.5
Dark
Green

A

YFP-SL-Tet

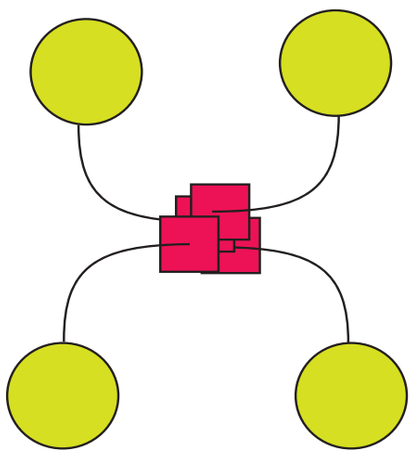
Tetramer@C

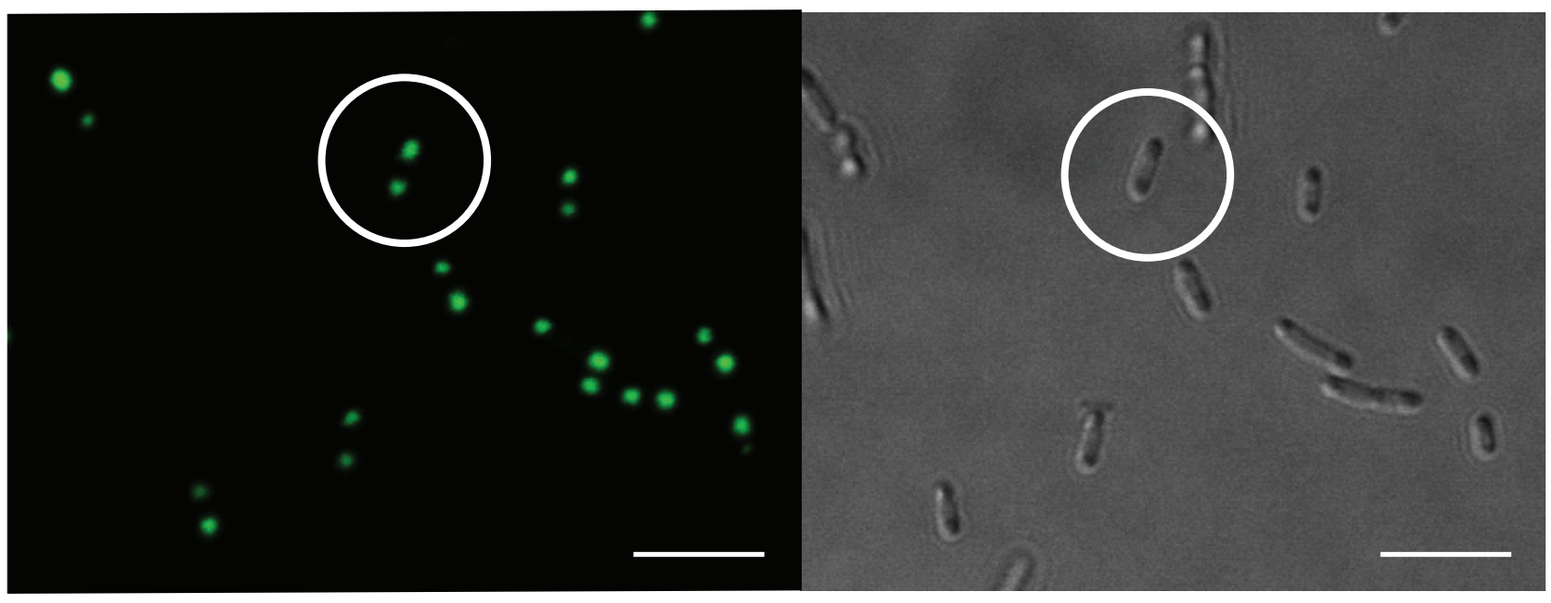




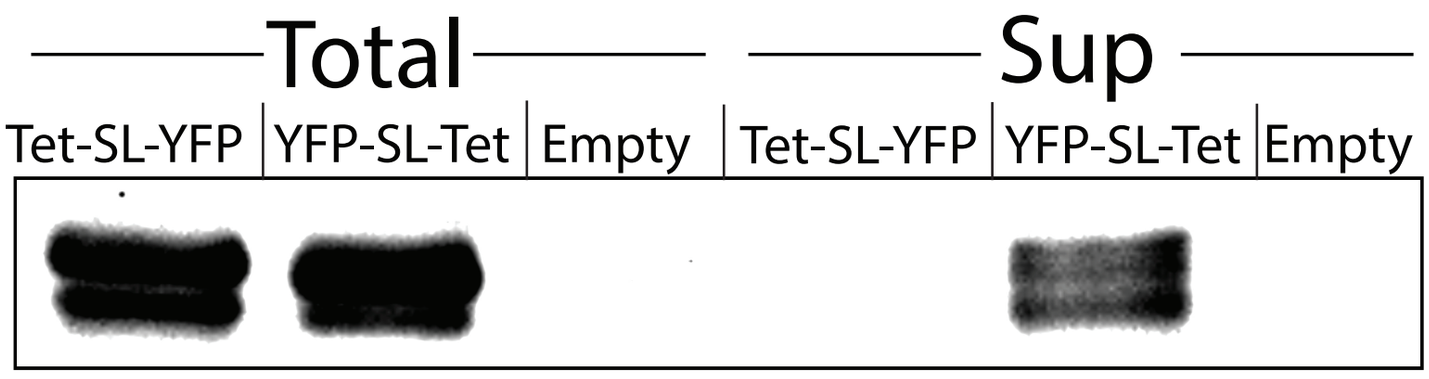
Tet-SL-YFP

Tetramer@N

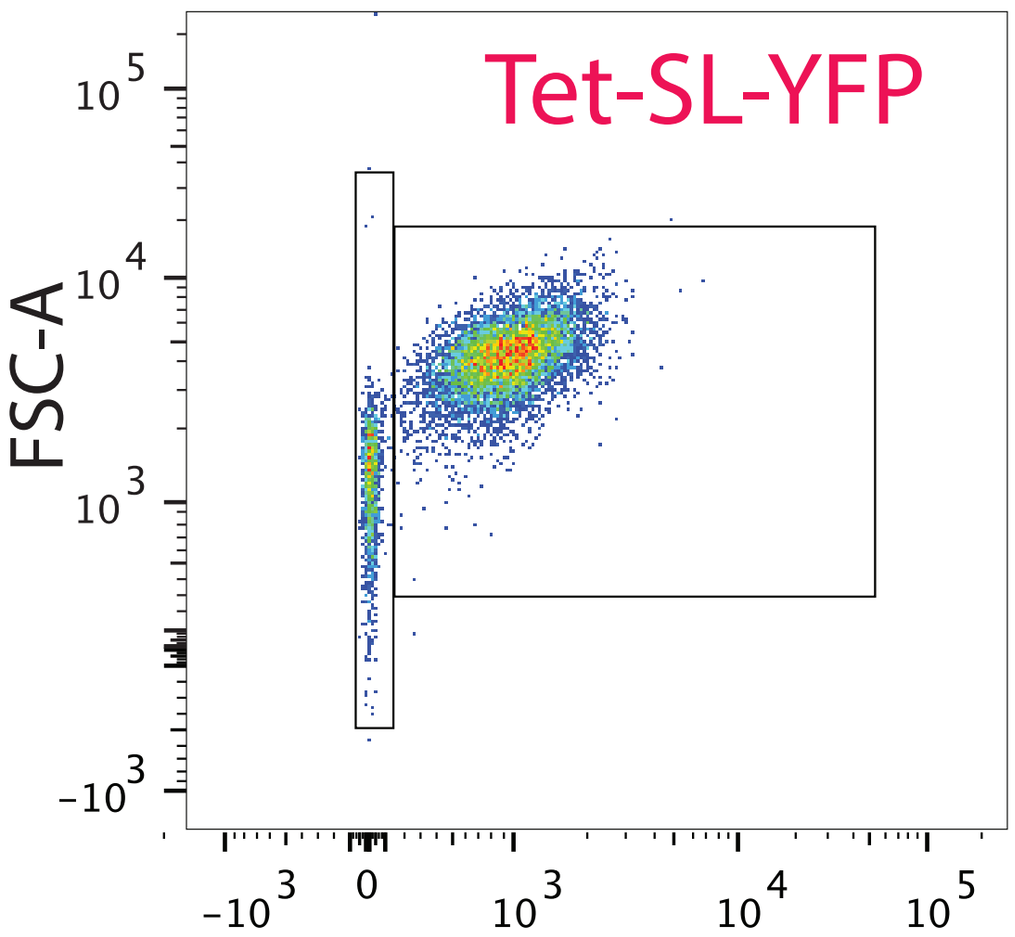


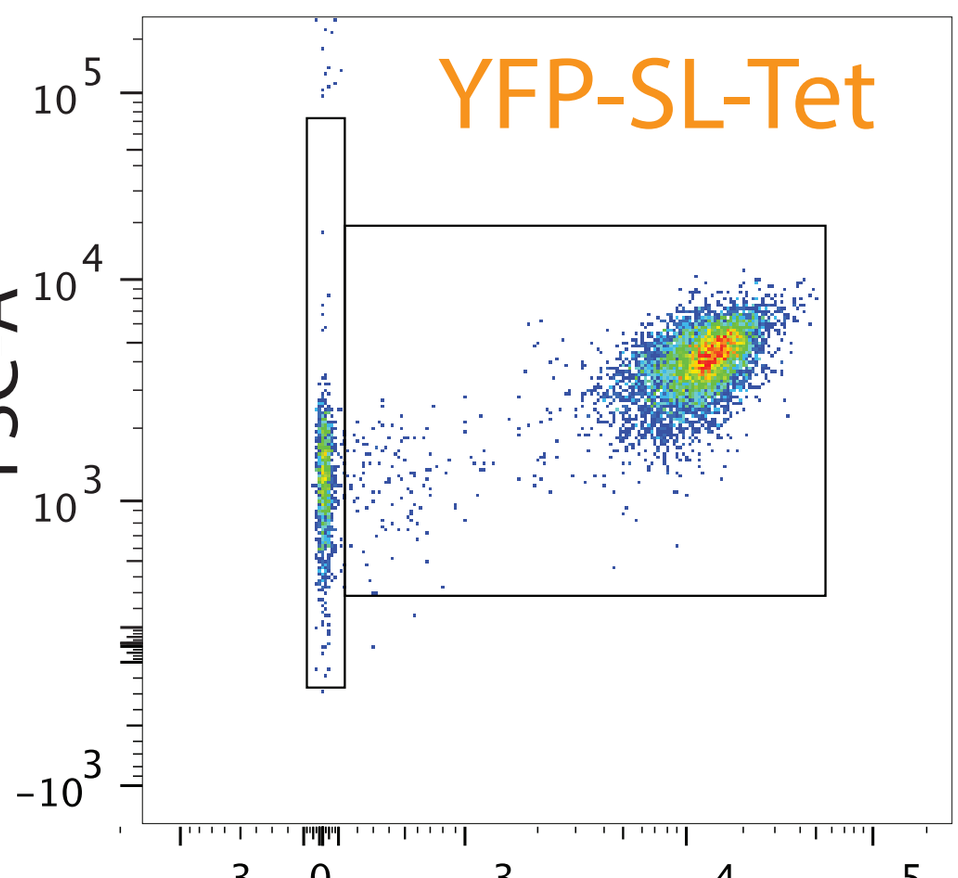


B



C

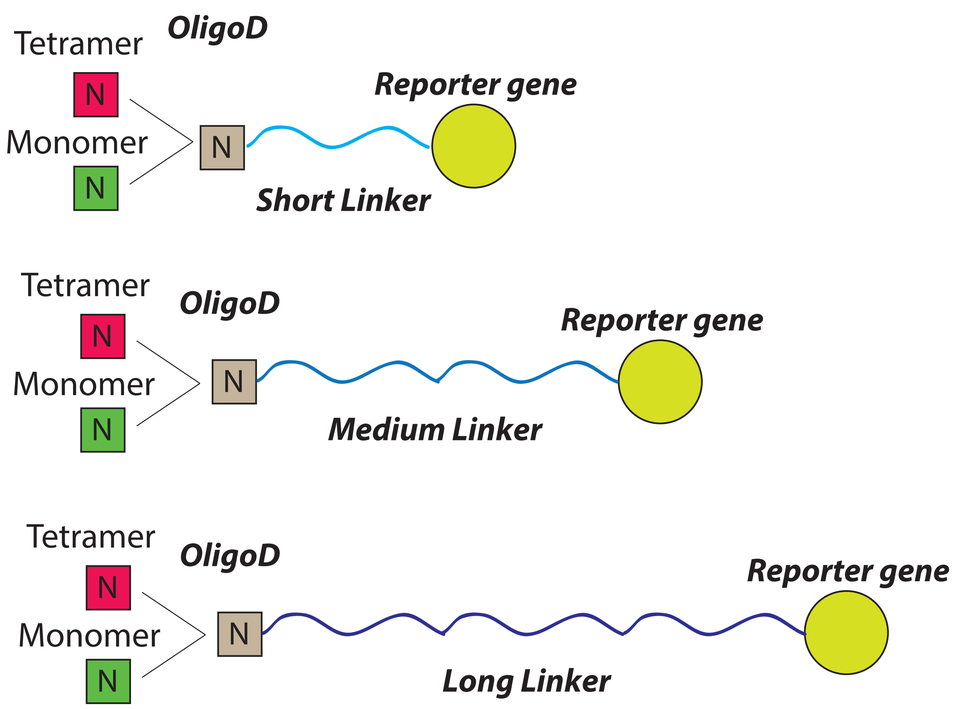




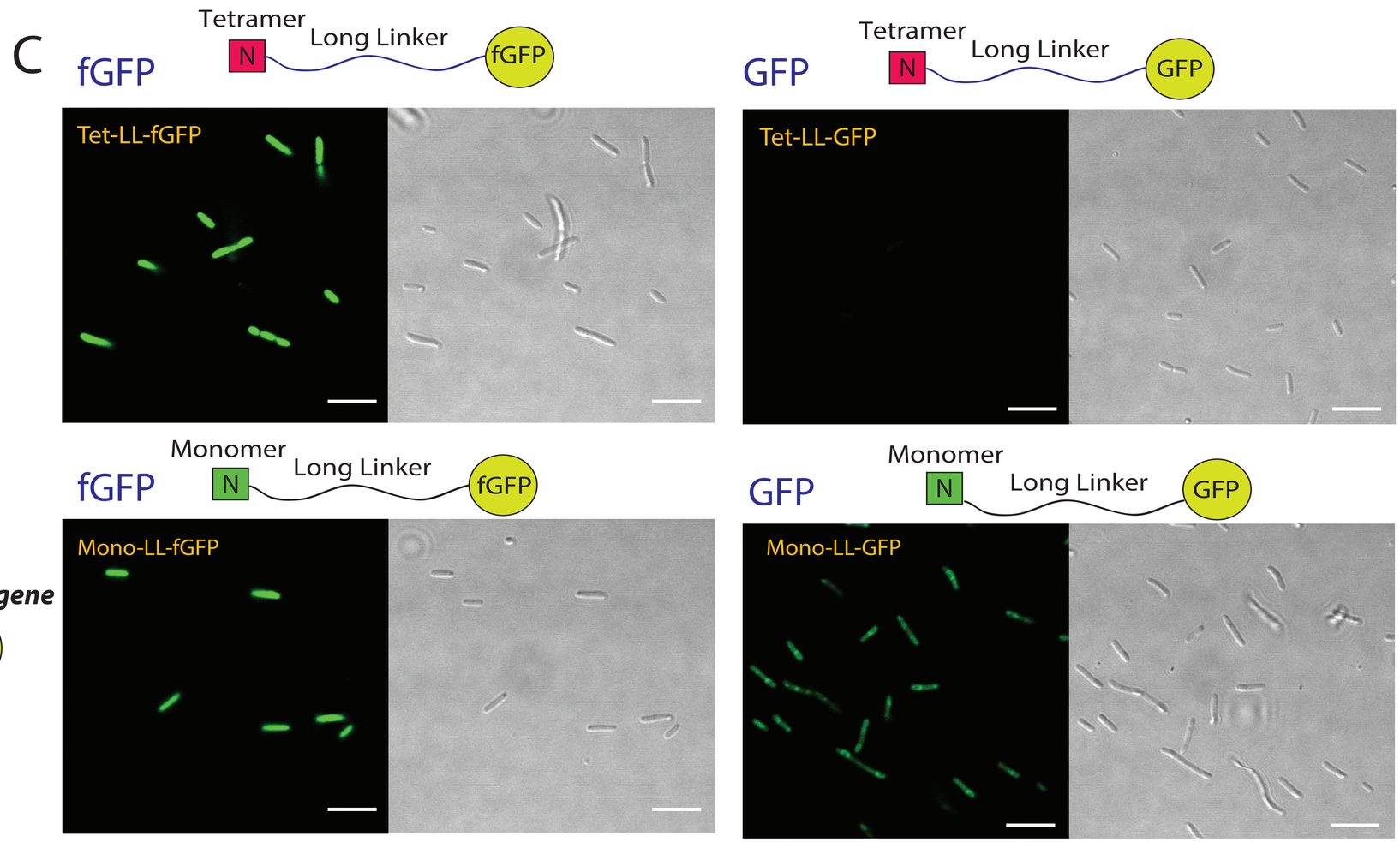
D

E

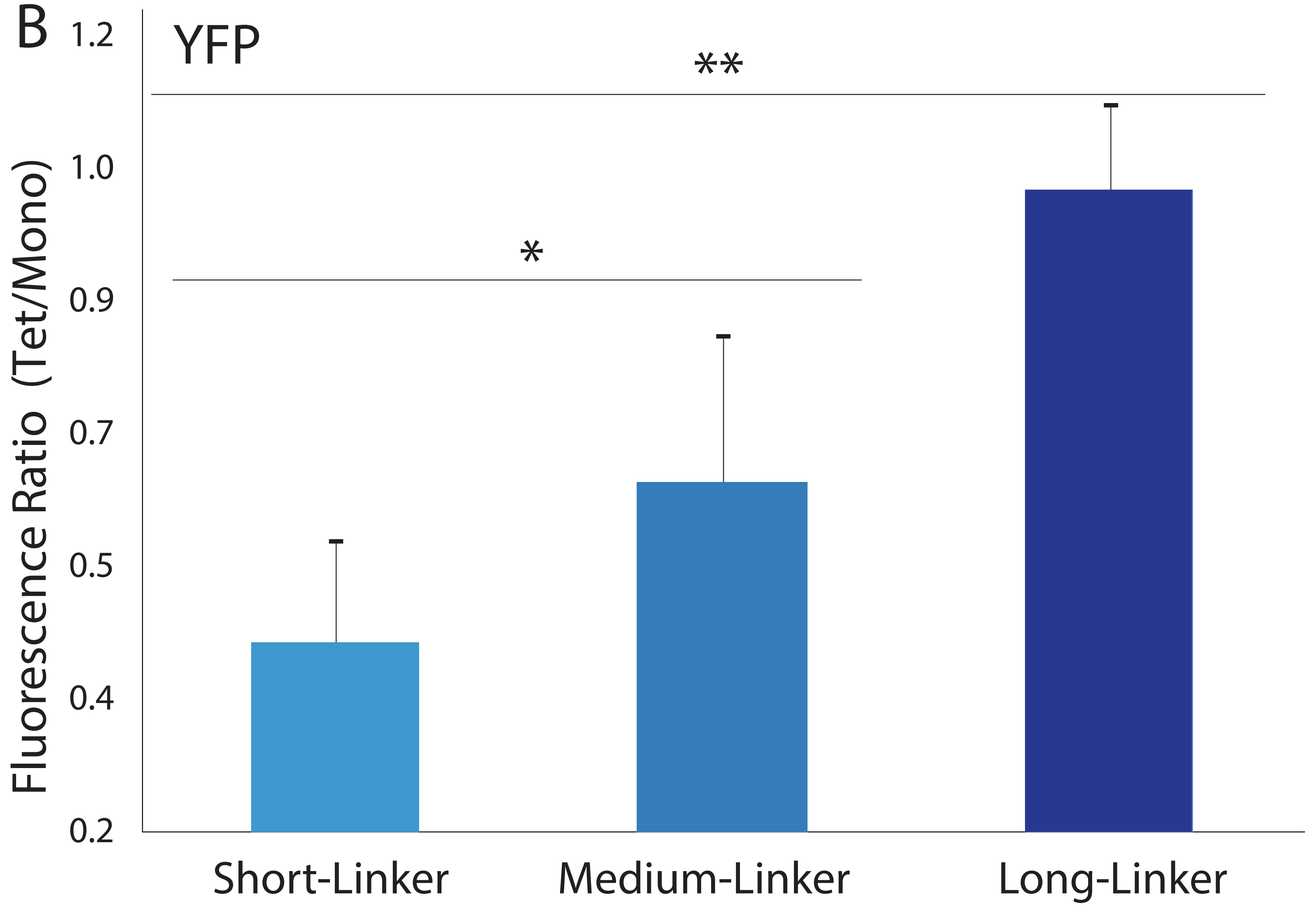
A


Tetramer
OligoD
N
Monomer
N
N
Reporter gene
Short Linker
Tetramer
OligoD
N
Monomer
N
N
Reporter gene
Medium Linker
Tetramer
OligoD
N
Monomer
N
N
Reporter gene
Long Linker

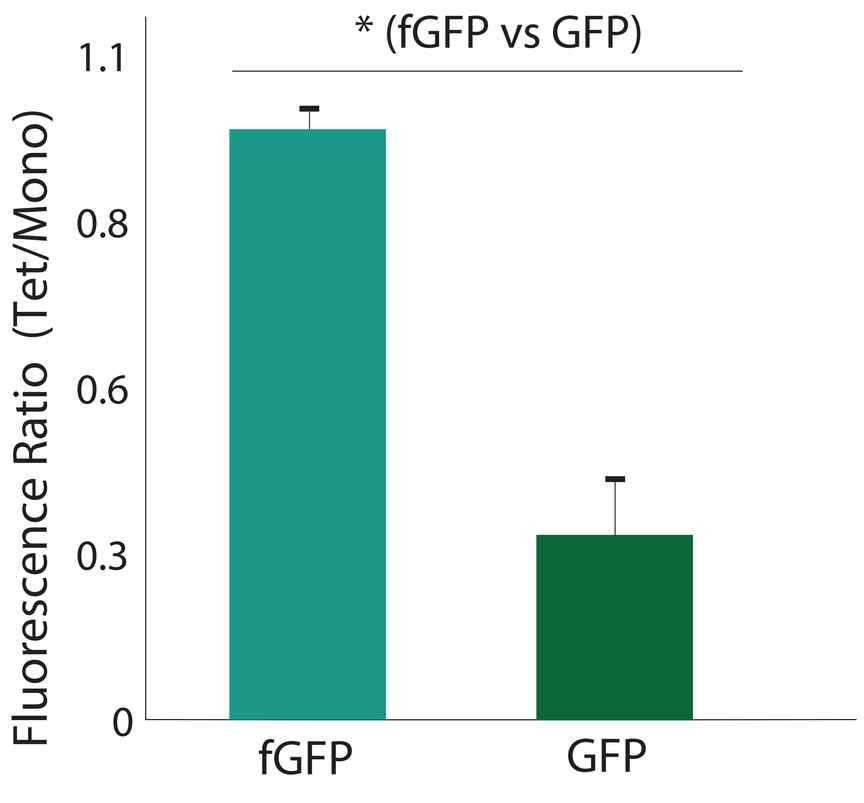
C


fGFP
Tetramer
N
Long Linker
fGFP
GFP
Tetramer
N
Long Linker
GFP
Tet-LL-fGFP
Tet-LL-GFP
fGFP
Monomer
N
Long Linker
fGFP
GFP
Monomer
N
Long Linker
GFP
Mono-LL-fGFP
Mono-LL-GFP

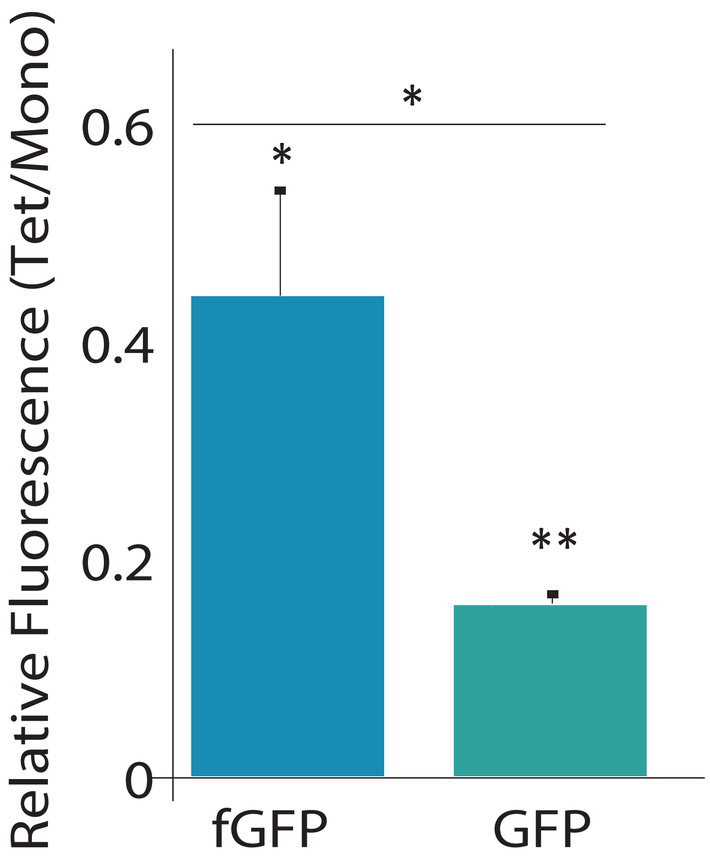
B


YFP
**
*
Fluorescence Ratio (Tet/Mono)
1.2
1.0
0.9
0.7
0.5
0.4
0.2
Short-Linker
Medium-Linker
Long-Linker

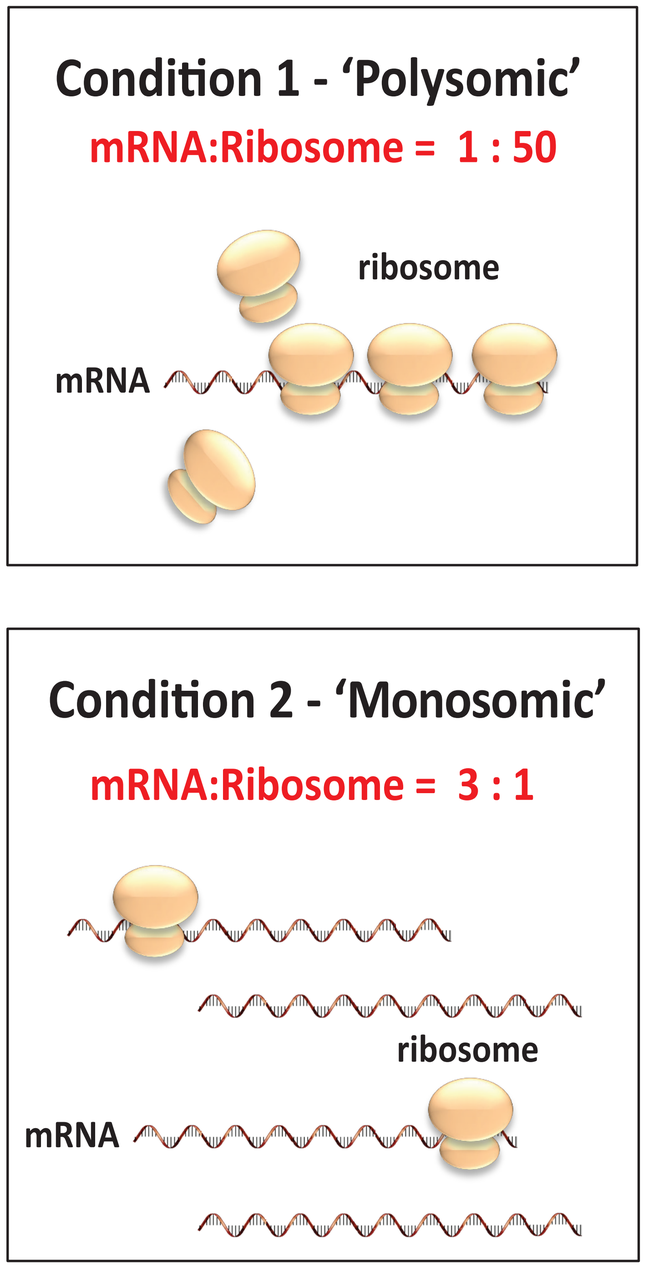
D

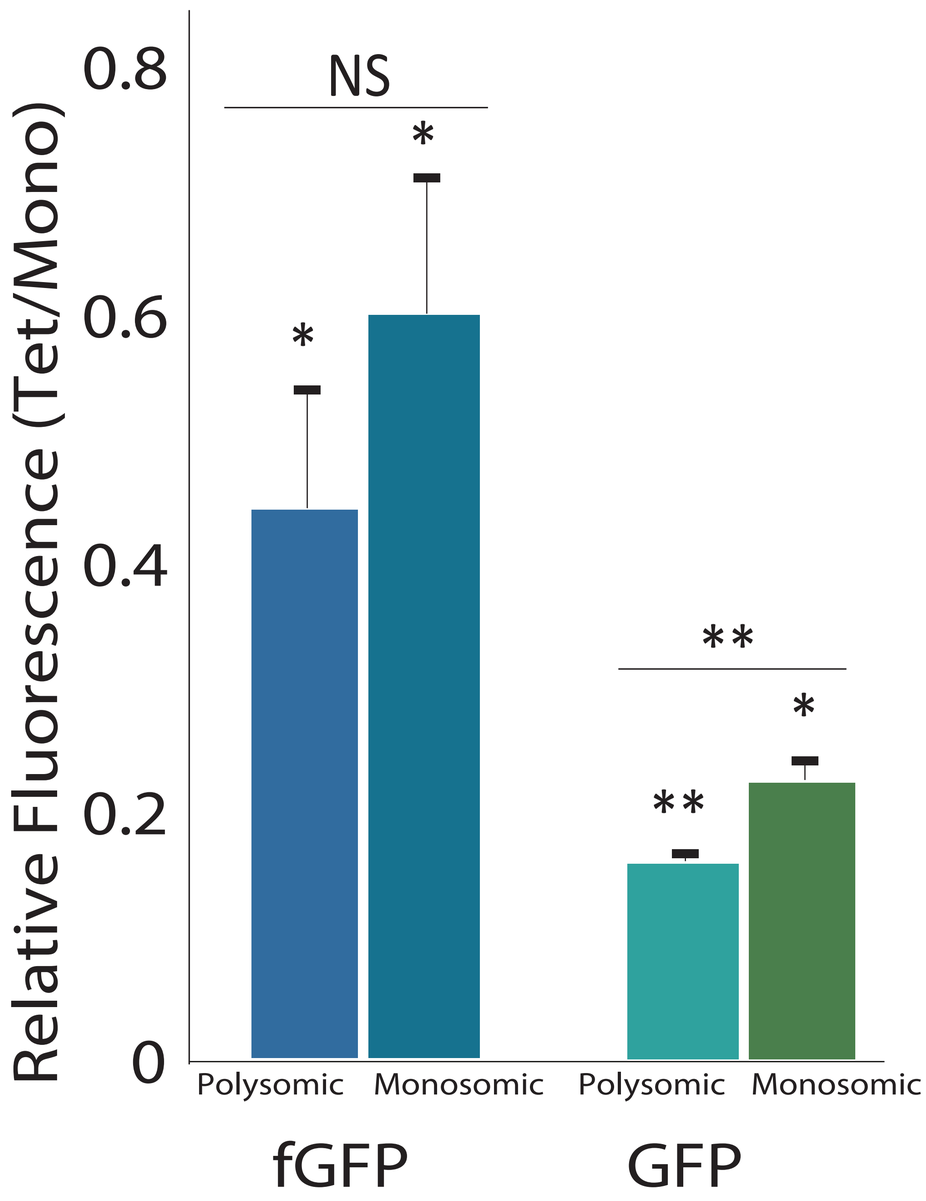

* (fGFP vs GFP)
Fluorescence Ratio (Tet/Mono)
1.1
0.8
0.6
0.3
0
fGFP
GFP

A

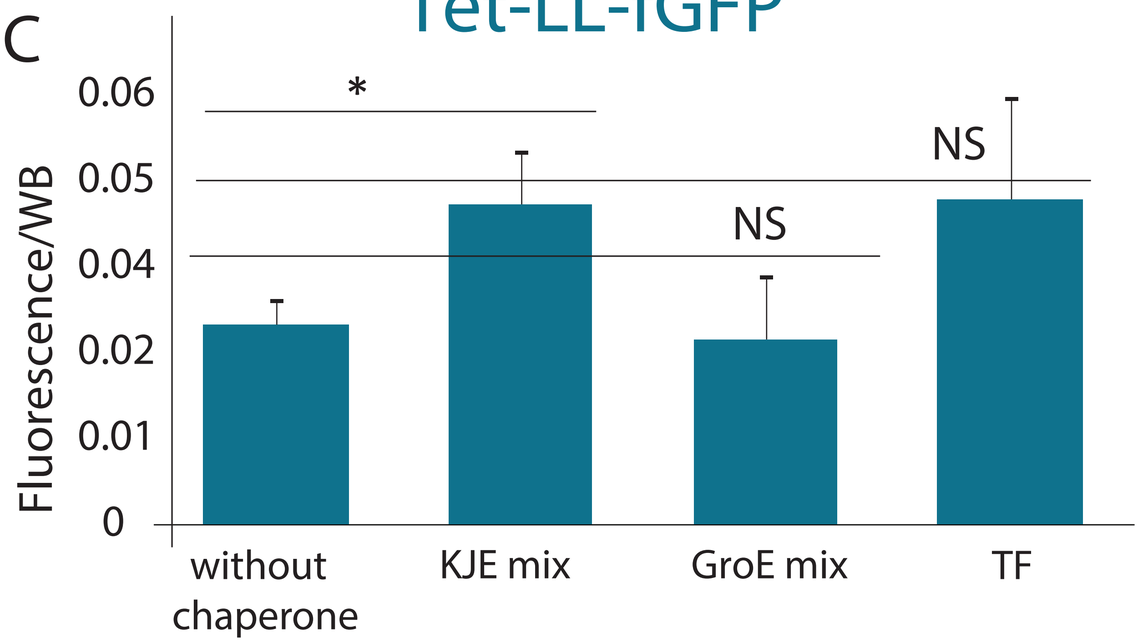

fGFP vs GFP
Relative Fluorescence (Tet/Mono)
0
0.2
0.4
0.6
*
*
**
fGFP
GFP

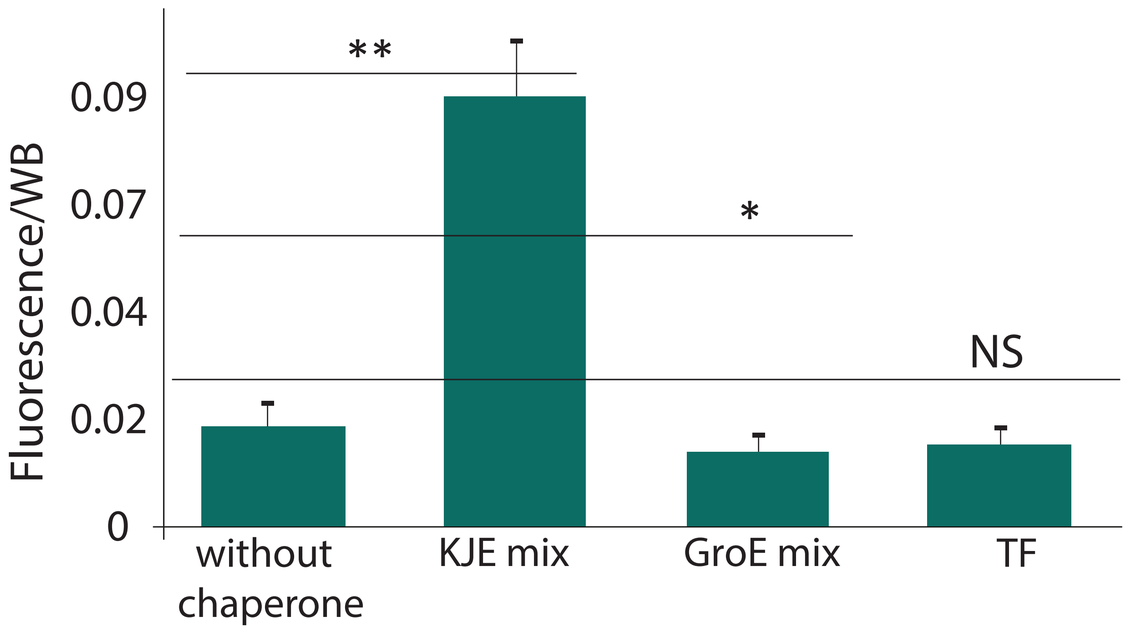
B


Condition 1 - 'Polysomic'
mRNA:Ribosome = 1 : 50
ribosome
mRNA
Condition 2 - 'Monosomic'
mRNA:Ribosome = 3 : 1
ribosome
mRNA

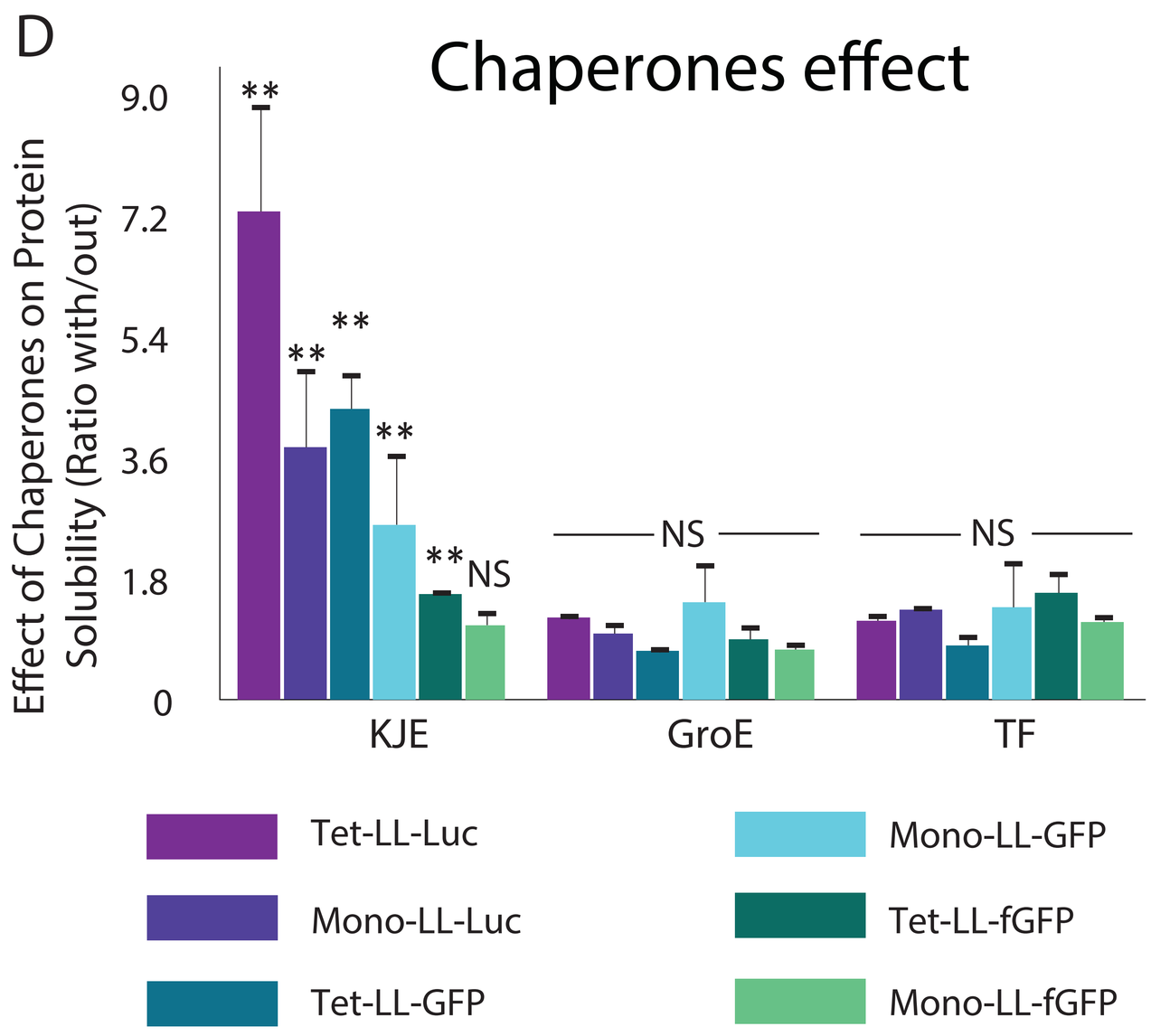

Polysomic vs Monosomic
Relative Fluorescence (Tet/Mono)
0
0.2
0.4
0.6
0.8
NS
*
*
**
**
*
Polysomic
Monosomic
Polysomic
Monosomic
fGFP
GFP

C

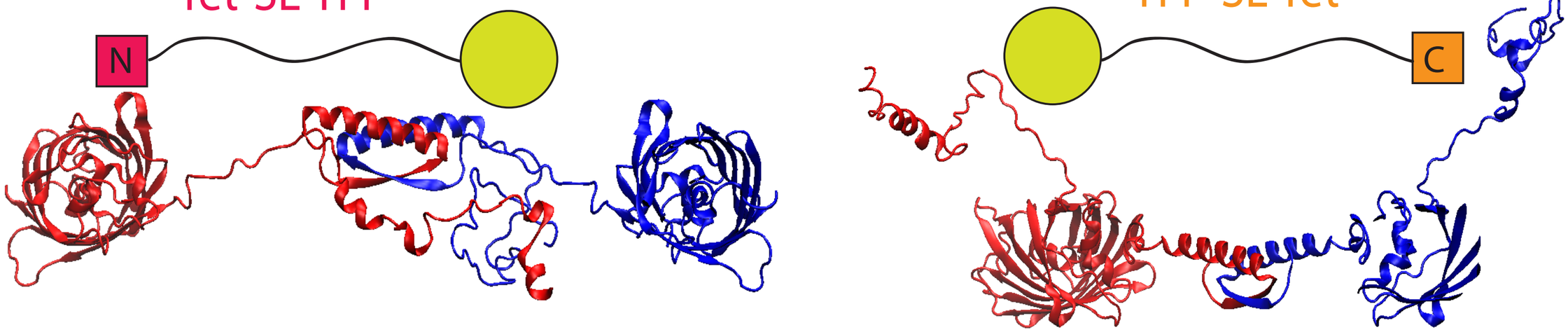

Tet-LL-fGFP
Fluorescence/WB
0
0.01
0.02
0.04
0.05
0.06
*
NS
NS
without chaperone
KJE mix
GroE mix
TF


Tet-LL-GFP
Fluorescence/WB
0
0.02
0.04
0.07
0.09
**
*
NS
without chaperone
KJE mix
GroE mix
TF

D

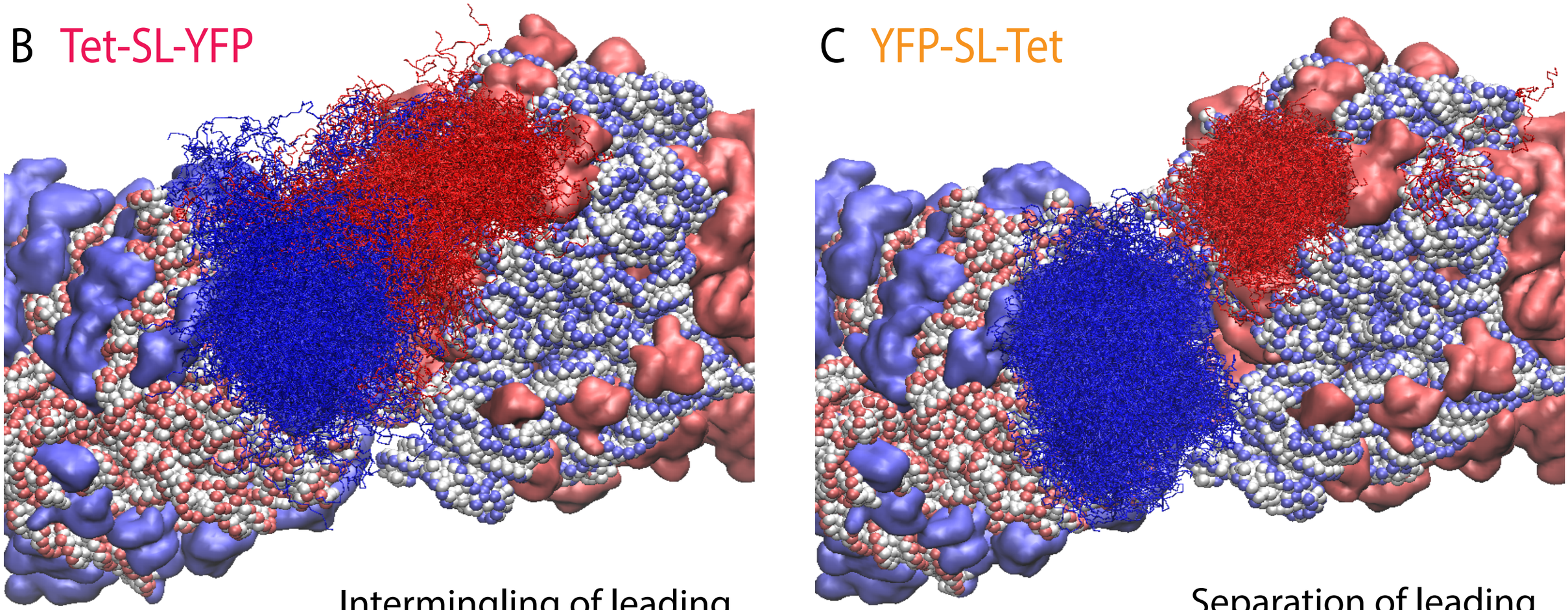

Chaperones effect
Effect of Chaperones on Protein Solubility (Ratio with/out)
0
1.8
3.6
5.4
7.2
9.0
**
**
**
**
NS
NS
NS
KJE
GroE
TF
Tet-LL-Luc
Mono-LL-Luc
Tet-LL-GFP
Mono-LL-GFP
Tet-LL-fGFP
Mono-LL-fGFP

A



B Tet-SL-YFP

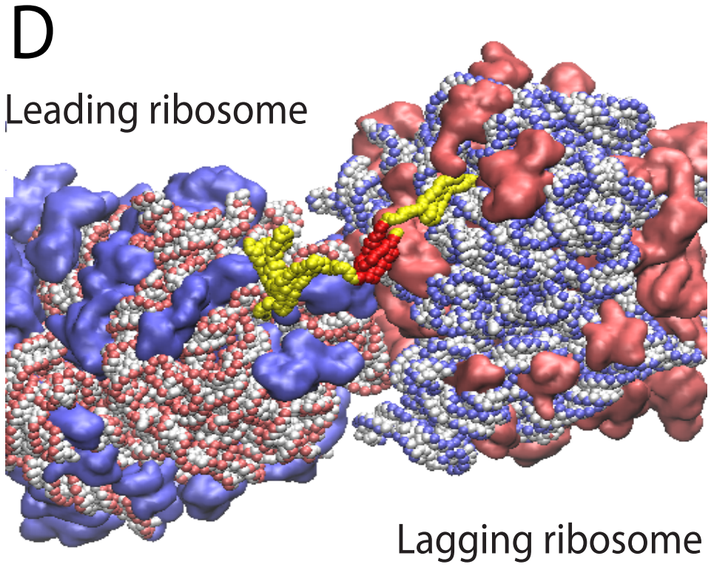
C YFP-SL-Tet



Intermingling of leading and lagging chains

Separation of leading and lagging chains

D

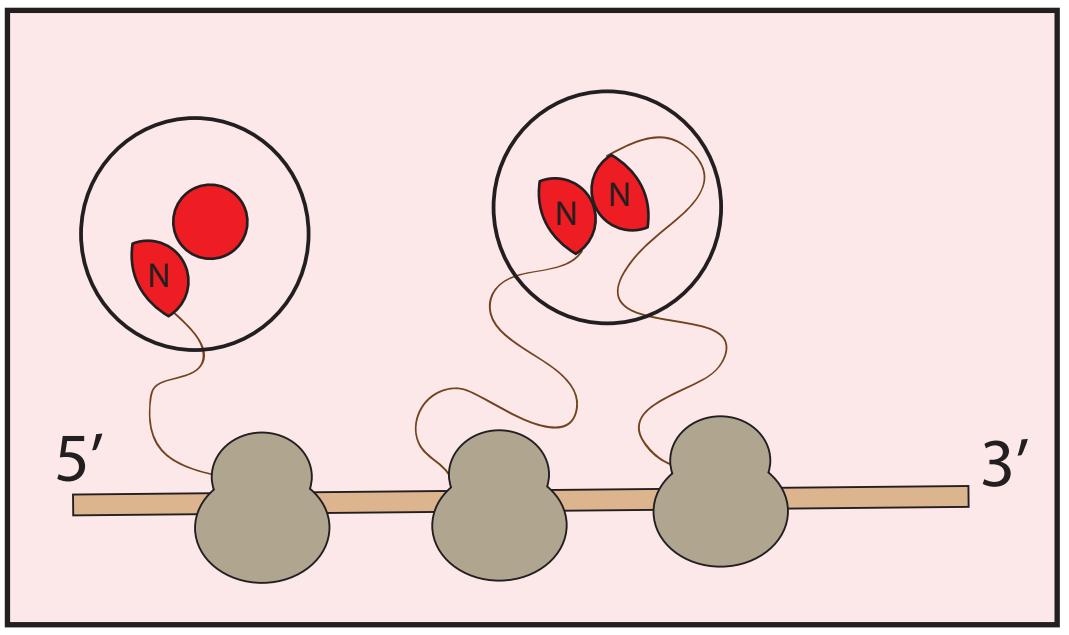


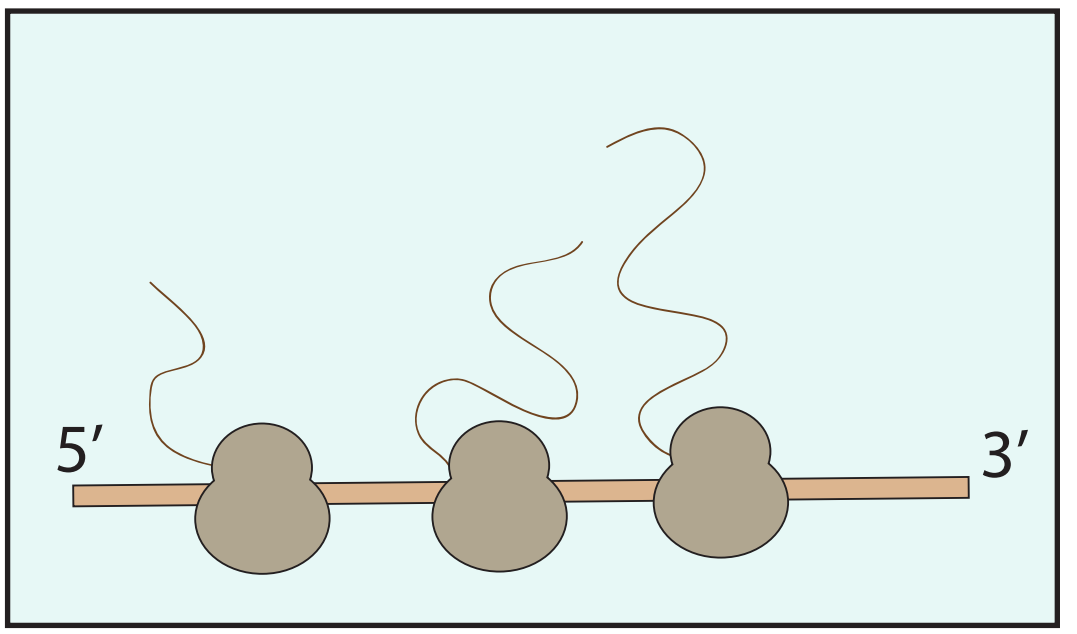
E

	Tet-SL-YFP	Tet-LL-YFP	YFP-SL-Tet
Co-translational Assembly	18	15	0 (3)
No Assembly	2	5	20 (17)
Misassembly-like	15	3	0 (4)
Total Simulation	20	20	20

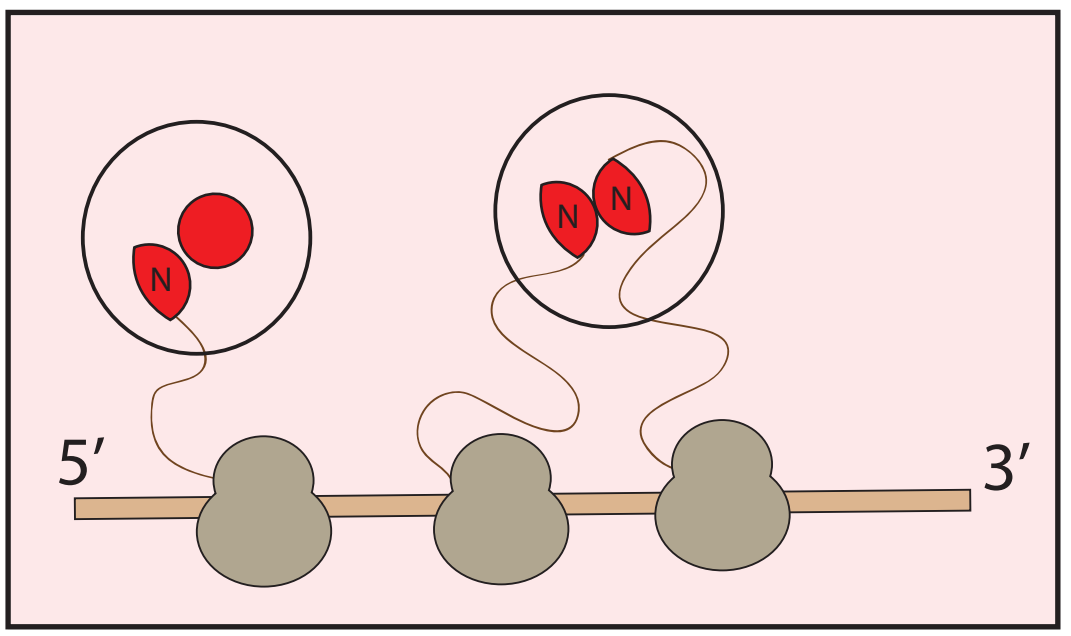
A

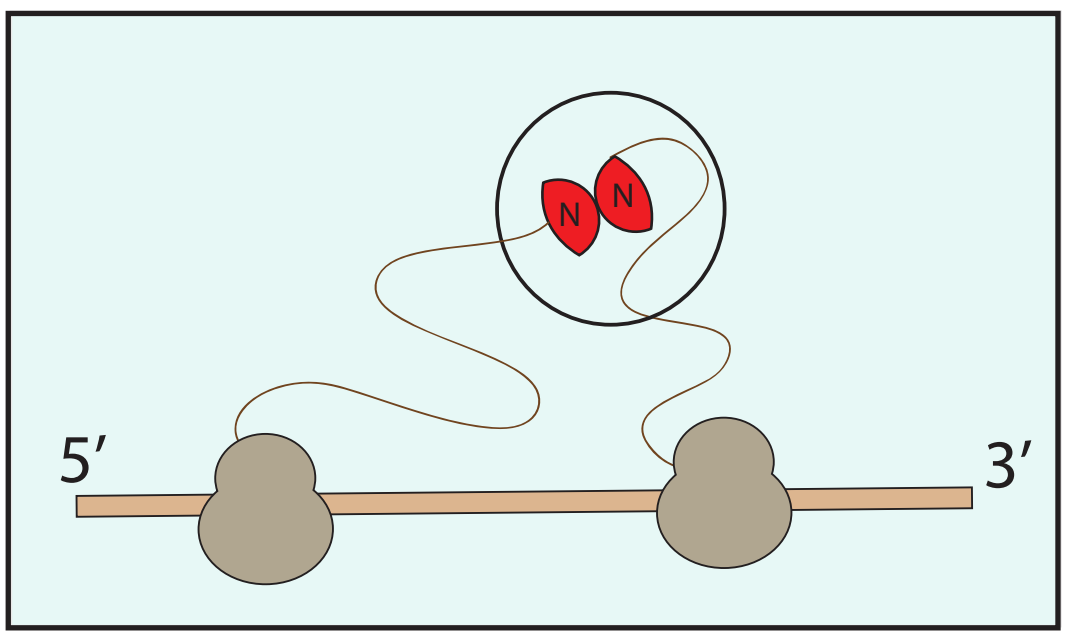
N- vs. C-oligomerization Domain



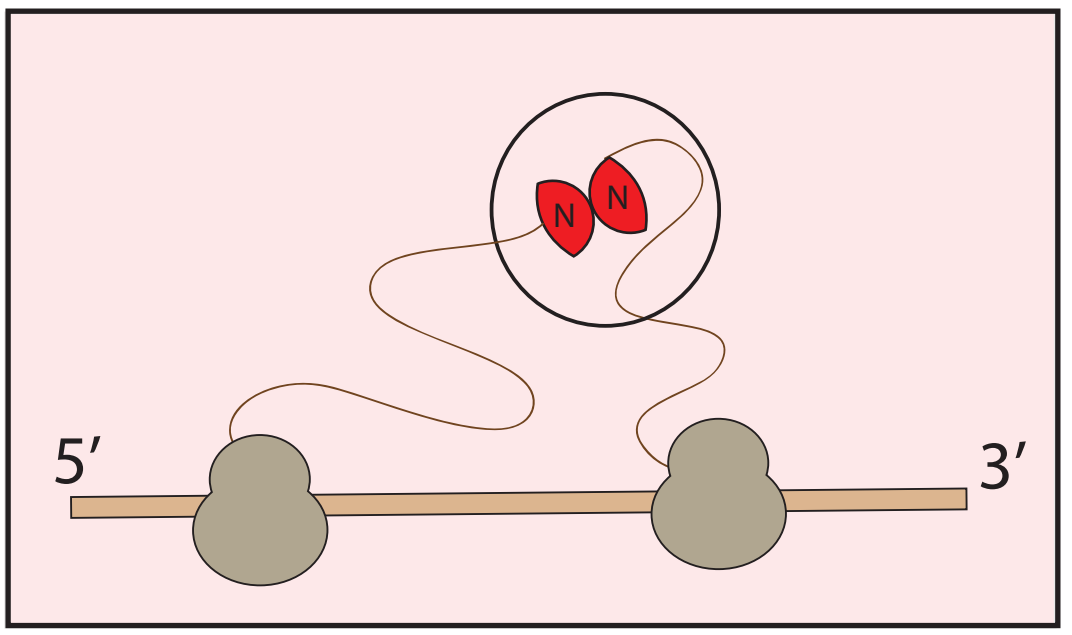


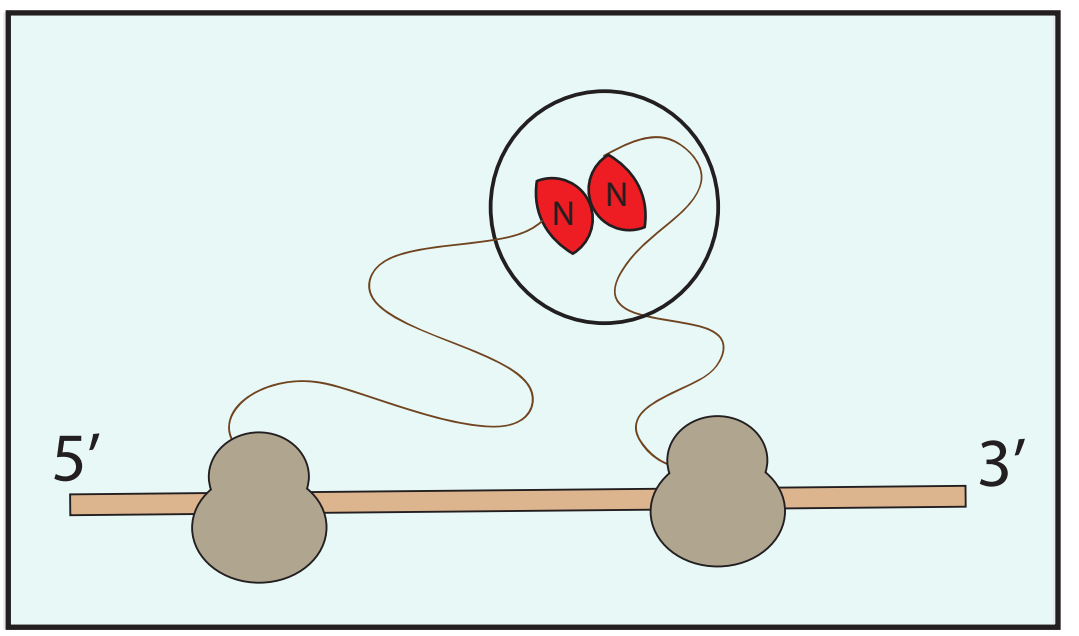
Short vs. Long Linker





Slow vs. Fast Folding-Rate





B

’High Risk’	’Low Risk’
Assembly via N	Assembly via C
Short Linker	Long Linker
Slow Folding	Fast Folding
Polysome	Monosome
Cotrans-folding	Posttrans-folding

C